Unsupervised Model Adaptation for Continuous Speech Recognition Using Model-Level Confidence Measures

KWAN Ka Yan

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy
in
Electronic Engineering

June 2002

Abstract of thesis entitled:

Unsupervised Model Adaptation for Continuous Speech Recognition Using Model-Level Confidence Measures

Submitted by **KWAN Ka Yan**

for the degree of **Master of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in **June 2002.**

Model adaptation aims at improving the performance of automatic speech recognition systems by reducing the acoustic mismatch between training data and the input utterance to be recognized. In unsupervised adaptation, the transcription of adaptation utterances is not known. Text output generated by the recognizer is used instead. Given that no recognizer is perfect, it is inevitable that erroneous information is used for adaptation. This may lead to incorrect adjustment of the recognition system. To alleviate such effect, confidence measure is applied to identify the adaptation data that are recognized with low reliability.

Confidence measures have been widely used for the detection of speech recognition errors. They are also useful in selecting and screening data for unsupervised adaptation of hidden Markov model (HMM). In our work, a model-level confidence measure is

proposed for model adaptation with the Maximum Likelihood Linear Regression (MLLR) techniques. Model-level confidence measure provides a finer selection of adaptation data than word or utterance level measures. Moreover, we propose to compute confidence score based on not only the recognized models but also other models that are easily confused with them are involved in the computation.

The proposed methods are evaluated in three recognition tasks that are different in recognition domain and/or acoustic channels. It is found that these two conditions affect the performance of confidence measure significantly. Experimental results show that the proposed confidence measures improve the effectiveness of unsupervised model adaptation. The relative improvements in word error rate are up to 9.75%.

摘要

模型自適應技術能藉著減低訓練與識別環境間的不匹配，有效地提高語音識別系統的準確性。在無監督的算法框架下，識別的誤差直接影響自適應效果。爲了篩選合適的數據，我們應用了識別結果可信度作爲依據。

可信度被廣泛應用於監測識別上的失誤，它也被應用於非監督性的馬爾可夫模型 (hidden Markov model) 自適應算法中的數據選取。我們爲自適應性技術—最大似然線性回歸 (Maximum Likelihood Linear Regression) 提出了模型層次上的可信度。相對詞語或句子的可信度，模型層次上的可信度提供了更細緻的數據篩選。我們提出的另一個算法不單用了被識別的模型，也用上了容易被混淆的模型對可信度行估算。

我們在不同的語音環境及範疇進行了實驗，發現了這兩個因素對可信度的應用成效有很大影響。實驗結果証明了我們的方法能有效地改善非監督的自適應結果，字辨識別差率能減少了 9.75%。

Acknowledgment

I would like to express my sincere gratitude to Prof. Tan Lee for his supervision and insightful advice throughout this research. The valuable suggestions and support from Prof. P. C. Ching are much appreciated. Thanks are due to W. N. Choi, who has given many useful advices and support. I would like to thank Y. W. Wong for his guidance and valuable suggestion. Thanks are also due to W. K. Lo and W. Lau for their technical assistance.

I would like to thank all the colleagues and friends in DSP group. They have helped me in many different ways. Thanks are given to S. K. Cheung, Patgi Kam, Arthur Luk, K. M. Law, W. Lam, L. Y. Ngan, Herman Yau, K. F. To and S. W. Wu. Special thanks are given to those who help me to proofread my thesis. It was a pretty difficult job.

Finally, I would like to express my sincere gratitude to my parents and my sister for their patience, support and understanding throughout this research.

Thanks for God to give me strength.

Contents

1. Introduction 1

1.1. Automatic Speech Recognition 1

1.2. Robustness of ASR Systems 3

1.3. Model Adaptation for Robust ASR 4

1.4. Thesis outline 6

References 8

2. Fundamentals of Continuous Speech Recognition 10

2.1. Acoustic Front-End 10

2.2. Recognition Module 11

2.2.1. Acoustic Modeling with HMM 12

2.2.2. Basic Phonology of Cantonese 14

2.2.3. Acoustic Modeling for Cantonese 15

2.2.4. Language Modeling 16

References 17

3. Unsupervised Model Adaptation 18

3.1. A General Review of Model Adaptation 18

3.1.1. Supervised and Unsupervised Adaptation 20

3.1.2. N-Best Adaptation 22

3.2. MAP 23

3.3. MLLR 25

3.3.1. Adaptation Approach 26

3.3.2. Estimation of MLLR regression matrices 27

3.3.3. Least Mean Squares Regression 29
3.3.4. Number of Transformations 30
3.4. Experiment Results 32
3.4.1. Standard MLLR versus LMS MLLR 36
3.4.2. Effect of the Number of Transformations 43
3.4.3. MAP Vs. MLLR 46
3.5. Conclusions 48
References xlix
4. Use of Confidence Measure for MLLR based Adaptation 50
4.1. Introduction to Confidence Measure 50
4.2. Confidence Measure Based on Word Density 51
4.3. Model-level confidence measure 53
4.4. Integrating Confusion Information into Confidence Measure 55
4.5. Adaptation Data Distributions in Different Confidence Measures 57
References 65
5. Experimental Results and Analysis 66
5.1. Supervised Adaptation 67
5.2. Cheated Confidence Measure 69
5.3. Confidence Measures of Different Levels 71
5.4. Incorporation of Confusion Matrix 81
5.5. Conclusions 83
6. Conclusions 85
6.1. Future Works 88

Lists of Figures

Figure 1-1: The flow diagram of an ASR system 2

Figure 1-2: The flow diagram of model adaptation 4

Figure 2-1: Topology of a HMM 13

Figure 3-1: The flowchart of supervised adaptation 20

Figure 3-2: The flowchart of unsupervised adaptation 21

Figure 3-3: The geometric description of MAP 24

Figure 3-4: A binary regression class tree 31

Figure 3-5: An example of a binary regression class tree 32

Figure 3-6: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformation in Task 1. 37

Figure 3-7: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformation in Task 2. 38

Figure 3-8: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformations in Task 1 (40 adaptation sentences). 40

Figure 3-9: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformation in Task 2 (40 adaptation sentences). 41

Figure 3-10: The WER (%) after applying supervised adaptation and unsupervised adaptation using N-best hypotheses in Task 1. 43

Figure 3-11: The WER (%) after applying supervised adaptation and unsupervised adaptation using N-best hypotheses in Task 2. 44

Figure 3-12: The WER (%) after applying supervised adaptation and unsupervised adaptation using N-best hypotheses in Task 3. 46
Figure 3-13: The WER(%) of MAP with 20 adaptation sentences 47
Figure 3-14: The WER(%) of MLLR with 20 adaptation sentences 47
Figure 4-1: Word graph of the 4-best hypotheses 52
Figure 4-2: Example of the incorporation of confusion matrix 56
Figure 4-3: (a) The amount of correctly recognized data for which the confidence scores are above or equal to difference thresholds in Task 1. (b) The amount of incorrectly recognized data for which the confidence scores are above or equal to the thresholds in Task 1 59
Figure 4-4: (a) The amount of correctly recognized data above or equal to the threshold of confidence measure in Task 2. (b) The amount of incorrectly recognized data above or equal to the threshold of confidence measure in Task 2 61
Figure 4-5: (a) The amount of correctly recognized data above or equal to the threshold of confidence measure in Task 3. (b) The amount of incorrectly recognized data above or equal to the threshold of confidence measure in Task 3 63
Figure 5-1: The WER (%) in Task 1 when confidence measures of different levels are integrated. "word-level", "biphone-based" and "baseIF-based" denote different levels of confidence measure. "no CM" denotes adaptation without using confidence measure. "supervised" demotes supervised adaptation. "cheated" denotes adaptation with cheated confidence measure. 71
Figure 5-2: The percentage of data above or equal to the threshold of confidence measure so they are selected for adaptation 72

Figure 5-3: The ratio of correctly recognized data in the selected data (Quality of the selected data) 72
Figure 5-4: The amount of correctly recognized data 73
Figure 5-5: The distribution of incorrectly recognized data which are in the same regression class with the actual transcription alignment and those which are in the other classes. 74
Figure 5-6: The WER (%) in Task 2 when confidence measures of different levels are integrated. 75
Figure 5-7: The Quantity of data which is represented by the amount of data 75
Figure 5-8: The quality of data which is determined by the ratio of correctly recognized data in selected data 76
Figure 5-9: The amount of selected correctly recognized data out of all correctly recognized data 76
Figure 5-10: The WER (%) in Task 3 when confidence measures of different levels are integrated 78
Figure 5-11: The quantity of data is determined by the percentage of data above the threshold 78
Figure 5-12: The quality of data which is determined by the ratio of correctly recognized data in selected data with different threshold 79
Figure 5-13: The percentage of correctly recognized data above or equal to the threshold 79
Figure 5-14: The WER(%) of biphone-based confidence measure incorporating with confusion matrix in Task 1 81
Figure 5-15: The WER(%) of biphone-based confidence measure incorporating with confusion matrix in Task 2 82
Figure 5-16: The WER(%) of biphone-based confidence measure incorporating with

confusion matrix in Task 3 82

Lists of Tables

Table 2-1: Phonologies hierarchy of Cantonese Syllable 15

Table 2-2: Context-independent and biphone models sequence for the word sequence < Sil 兩手和黃 Sil> 15

Table 3-1: The RTF per sentence of different technique in Task 1. 37

Table 3-2: The RTF per sentence of different techniques in Task 2. 38

Table 3-3: The RTF of different technique in Task 1 (40 adaptation sentences). . 40

Table 3-4: The RTF per sentence of different technique in Task 2 (40 adaptation sentences). 41

Table 4-1: Part of the confusion matrix for biphone "*I_g+F_ai*" 56

Table 5-1: The WER(%) and relative improvement(%) in three tasks. 67

Table 5-2: The ratio of correct state and class in the model-level incorrectly recognized data in three tasks 70

Chapter 1

Introduction

Human-computer communication is highly concerned when computers are becoming more and more powerful and popular. We would like to see that such a communication system is just as effective as human-to-human communication. The first thing it should be able to do is that the computer understand what we are speaking. This is made possible by speech recognition techniques, which convert speech signals into text information.

1.1. Automatic Speech Recognition

Automatic speech recognition (ASR) [1] is the process that translates a piece of speech into text. In order to let the computer understand human speech, the mappings between acoustic speech signals and text symbols need to be established. The basic unit in this mapping is called acoustic model. An acoustic model may represent a phrase, word or phoneme depending on the application.

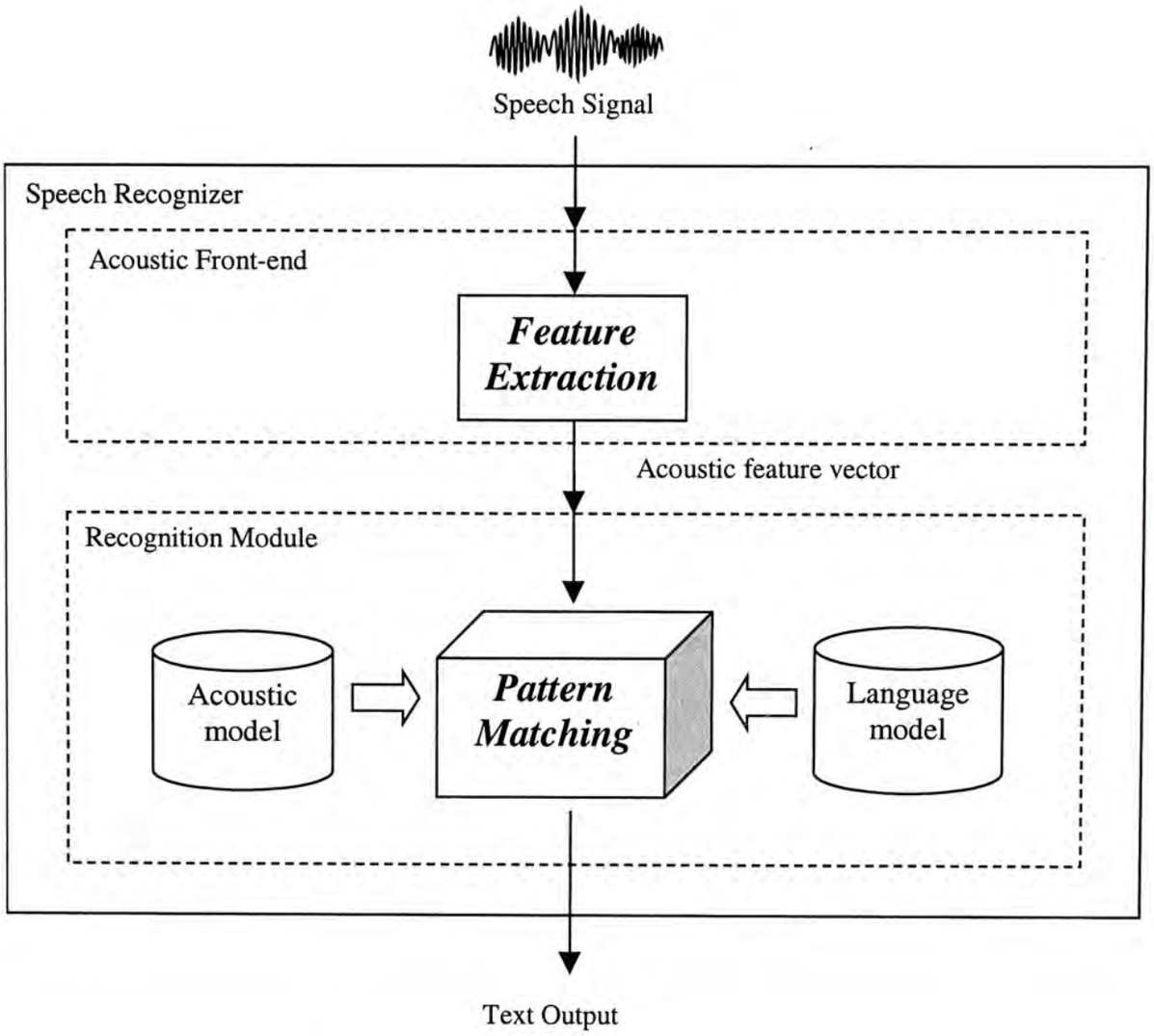


Figure 1-1: The flow diagram of an ASR system

Figure 1-1 shows the simple flow of an ASR system. The unknown utterance would be compared with all acoustic models in the database. Meanwhile, language models are used to impose linguistic constraints on the combination of words and the probability of word occurrence. These constraints help to exclude the output that is grammatically or lexically incorrect. In the pattern matching process, the best-matched acoustic model would be selected and the corresponding text symbols give the recognition result.

1.2. Robustness of ASR Systems

Usually, users of ASR system are unknown at the development stage. Therefore, the system should be robust to the variation of speakers and environment conditions. The acoustic models of such system are usually trained with utterances from different speakers or conditions so that the models can describe more general characteristics.

However, the generality is limited by the storage capacity of the acoustic models and the number of training speakers available. The general acoustic models can hardly represent all particular and special features under different operating conditions. This would cause the mismatch between the training utterances of the acoustic model and the input utterances to be recognized. For example, an acoustic model represents the speech of native speakers in a quiet environment, but the input speech is spoken by a non-native speaker in a noisy place. Such mismatch tends to degrade the recognition performance. Although general acoustic models can represent the general characteristic of different conditions, it may not be suitable to capture some specific features in different cases. The performance would be affected by such mismatches.

Many methods has been proposed to tackle this problem, such as combination of noise and clean speech model [2], prediction and modeling of pronunciation variations [3] and adaptation of acoustic models [4]-[7], which is the focus of this research.

The combination of noise and clean speech model can be applied in noisy recognition environment. The pronunciation variation dictionary can help to solve the problems of pronunciation variation of the same word. It is useful in the recognition of non-native speech. When the recognition condition is fixed, model adaptation can solve both of these problems.

1.3. Model Adaptation for Robust ASR

Model adaptation attempts to use a small amount of data, e.g. a few minutes of speech, to adjust the acoustic models in an ASR system. This is shown in Figure 1-2. The general acoustic models are tuned based on the adaptation speech and its content, in the hope of reducing the mismatch between them.

Model adaptation is one of the most common approaches for speaker adaptation [4]-[7]. It has been widely applied to voice dictation systems and voice-enabled inquiry system [8]. There are many model adaptation techniques: speaker clustering [4], Maximum a Posterior (MAP) [5], Maximum Likelihood Linear Regression (MLLR) [6] and Eigenvoice [7].

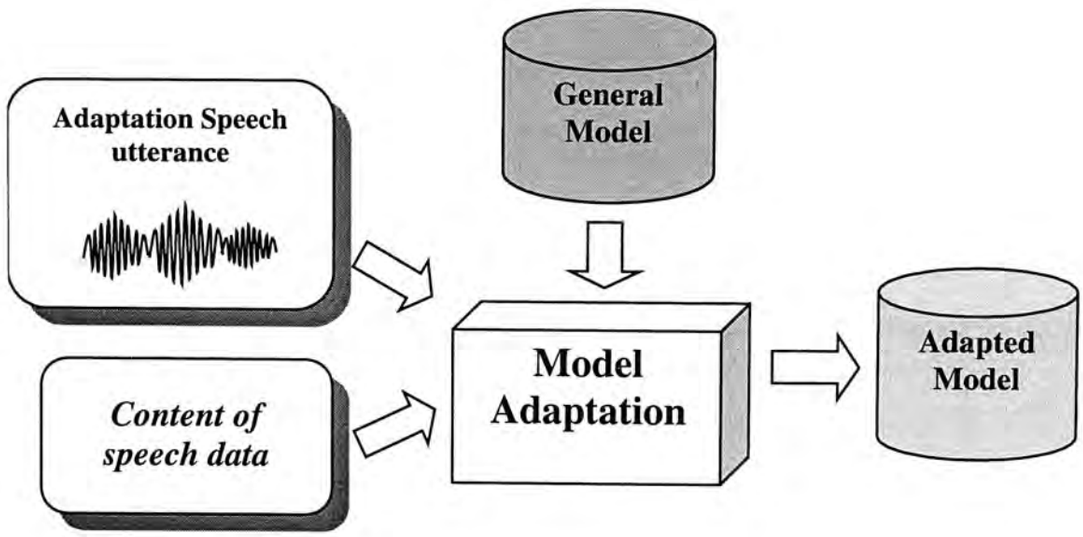


Figure 1-2: The flow diagram of model adaptation

Speaking clustering [4] is the simplest model adaptation technique to cluster the training speakers into different classes, may be according to their gender, speaking style or other speech characteristics. Instead of training only one general acoustic model, each class has its own acoustic model. The adaptation speech is firstly matched into a particular speaker/class and then the acoustic model of that class is used for recognition. Although the most suitable acoustic model for the input speech is selected, the parameter of acoustic model cannot be tuned by the input speech.

In order to improve performance of model adaptation, many techniques are proposed to adjust the acoustic model with the feature of input speech utterance. Eigenvoice [7] attempts to find the eigenvectors of the acoustic model. Based on the feature of adaptation speech, the newly adapted acoustic model would be formed as the combination of the eigenvectors. Maximum a Posterior (MAP) [5] tries to find the adapted model by using the linear combination of adaptation speech and acoustic models. Maximum Likelihood Linear Regression (MLLR) [6] is a transformation–based adaptation. The last two techniques will be discussed in Chapter 3.

Model adaptation can improve recognition performance. However, if the content of adaptation speech is not correct, the effectiveness of adaptation result would be affected. For example, in the adaptation speech, the speech data of word "一" is wrongly mapped to "七". Then the model of "七" is adapted by the speech data of "一". Therefore, when you speak "一" next time, it may be recognized as "七". To filter these errors in adaptation speech, "confidence measure" can be used.

Confidence measure attempts to represent the reliability of the recognition output [9] -[11]. There are many techniques of computing the confidence measure. One is using

the length of the recognized word. The longer the word, the more reliable the word is. Another approach is using the acoustic score or language model score of the recognized word to determine the confidence score. And there is a method to use the word lattice and summarize all the above methods [10]. The method we used in our work is called "the word density" [9]. The occurrence frequency of the word in the N-best hypotheses is used to determine the confidence measure. It is reasonable to believe a recognized word, that is found in most hypotheses, is correct. This method is chosen since it uses not only the information of acoustic and language model but also the occurrence frequency in the N-best hypotheses.

In order to facilitate model adaptation, we try to find the reliability of a model rather than a word. Besides, the confusion matrix is used in the computation of confidence score since the pronunciation variation degrades the recognition performance and affects the estimation accuracy of confidence measure. Adding the information of confusion matrix can further improve the confidence measure.

By the integration of confidence measure in model adaptation, the recognition performance is improved.

1.4. Thesis outline

The thesis will be divided into six chapters. A brief introduction of speech recognition system and a study of the acoustic model used in our project, hidden Markov model, will be given in the next chapter. Then we will discuss speaker adaptation in Chapter 3. MAP and MLLR will be compared and the adaptation performance in different tasks

will be discussed. In Chapter 4, we will talk about the word-density confidence measure and compare the word-level and model-level confidence measures. Moreover, the use of confusion matrix for modeling pronunciation variation and its incorporation into confidence measure will be described. The integration of confidence measure in model adaptation will also be evaluated in Chapter 5. Finally, conclusions will be given.

References

[1] L. Rabiner, B. H. Junag, "The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization", *Fundamentals of Speech Recognition*, Chpt. 2, 1993.

[2] M. J. F. Gales, "A Fast and Flexible Implementation of Parallel Model Combination", ICASSP-95, vol. 1, pp. 133-136, 1995.

[3] M. Liu, B. Xu, T. Hunng, Y. Deng, C. Li, "Mandarin accent adaptation based on context-independent /content-dependent pronunciation modeling", ICASSP'00, pp. 1025-1028, 2000.

[4] T. Kosoka, S. Sagayama, "Tree-Structured Speaker Clustering for Fast Speaker Adaptation", ICASSP 94', vol.1, pp.245-248, 1994.

[5] J. L. Gauvain, C. H. Lee, "Maximum a Posterior Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. SAP, Vol. 2, No. 2 , pp. 291 –298, April 1994.

[6] C. L. Leggetter, P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continues Density HMMs," Computer Speech and Language, No. 9, pp. 171-185, 1995.

[7] R. Kuhn, P. Nguyen, J.-C Junqua, L. Goldwasser, "Eigenfaces and Eigenvoices: Dimensionality reduction for specialized pattern recognition,", Multimedia Signall Processing, IEEE Second Workshop, pp.71-76, 1998.

[8] E. Thelen, "Long Term On-line Speaker Adaptation for Large Vocabulary Dictation", ICSLP 96', vol 4, pp.2139-2142, 1996.

[9] M. Weintraub, "LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting," ICASSP1995, pp. 297-300, 1995.

[10] F. Wessel, R. Schluter, K. Macherey, H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Volume: 9 Issue: 3, 2001.

[11] F. Wallhoff, D Willett, G. Rigoll, "Frame-Discriminative and Confidence-Driven Adaptation for LVSCR", ICASSP'00, pp. 1835-1838, 2000.

Chapter 2

Fundamentals of Continuous Speech Recognition

Automatic speech recognition (ASR) is a process to translate a piece of speech to text. The block diagram of an ASR system was given as in Figure 1-1. Input speech is converted to feature vectors in acoustic front-end. Then the feature vector is translated to a text output in recognition module.

2.1. Acoustic Front-End

In this process, the speech signal is converted into some parametric representations, commonly referred to as feature vectors. Feature vectors contain useful materials of the signal and must be compact in size. Usually, each feature vector is used to represent a short duration of speech (typically about 25 ms). The speech signal is assumed to be stationary over this duration.

Among many proposed feature vectors, Mel-Frequency Cepstral Coefficients (MFCC) [1], Linear Prediction Coefficients (LPC) [2], Perceptual Linear Predictive

(PLP) Coefficients [3] have been most commonly used for speech recognition. All of them try to extract the speech information according to the human perception of speech signal with a small number of parameters. All parameters are independent of each others.

In our work, MFCC is used as the acoustic feature, which is based on filterbank analysis. The mel-frequency scale provides a non-linear frequency resolution according to human perception. Feature vectors are generated every 10ms. Each vector represents 25 ms of the speech waveform. Each vector contains 39 elements with 12 MFCC and their energy, as well as their first and second derivatives.

2.2. Recognition Module

The Recognition Module is the core of the whole ASR process. It is a pattern recognition process in which we would like to find a word sequence W_{max} from all the possible W, by maximizing a posterior probability $P(W|O)$. Thus is the best match with a sequence of feature vector O.

$$W_{\max} = \arg\max_{W} P(W|O) \qquad \text{eq. (2-1)}$$

However, direct estimation of a posterior probability is not possible since there is a huge number of possible word sequences. Using the Bayes' Rule, eq. (2-1) can be written as

$$W_{\max} = \arg\max_{W} P(O|W)P(W) \qquad \text{eq. (2-2)}$$

$P(O|W)$ is the probability of O being observed when the word sequence W is given. $P(W)$ is provided by language model, which depends on the application and is independent of speaker and speaking environment. And, $P(O|W)$ is provided by acoustic model.

The acoustic model attempts to characterize the statistical variation of acoustic observations. The performance of acoustic modeling would be affected significantly when there exists mismatch between the input utterance and the speech data used to train the acoustic model. Model adaptation tries to adjust the acoustic model in order to reduce such mismatch.

The details of acoustic modeling using hidden Markov model (HMM) will be given below. In addition, since the recognition tasks being investigated in this research focus on Cantonese, the characteristics of this Chinese Dialect will be described.

2.2.1. Acoustic Modeling with HMM

An acoustic model represents a speech unit. The speech unit can be a phrase, a word or a phoneme. For a large vocabulary recognition system, it is impractical to model word or phrase level units since the number of required models would become uncontrollably large. Subword speech units, e.g. phonemes or phone-like units, are used. And a lexicon is used to describe each word by the combination of these subword units.

Hidden Markov model (HMM) has been commonly used for acoustic modeling [4]. It is a finite state machine that is suitable to model time-varying pattern like speech signal. The typical topology of an HMM for ASR is shown below:

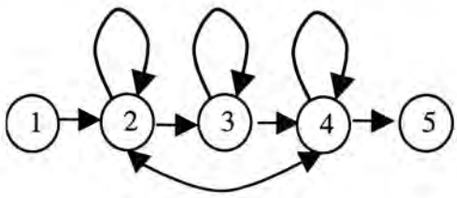


Figure 2-1: Topology of a HMM

Each HMM has three probability measures: initial state distribution probability, state transition probability and observation probability density function. The state transition probability is used to describe and control the state transition while the observation probability density function is used to model the statistical distribution at each state.

The beginning and ending states of an HMM do not contain any statistical parameter. It is used for the concatenation with other models. The transition between states is usually from left to right or self-looped. However, for some special speech units, such as silence, right-to-left or skip-state transitions may also be allowed. The state transition probability is equal to zero for illegal transitions and the rest are estimated in statistical approach.

Each state is represented as mixtures of multivariate Gaussian distributions. Each Gaussian distribution is specified by a mean parameter μ and a covariance matrix Σ. The components of a feature vector are assumed to be independent, so that Σ becomes

a diagonal matrix. By using diagonal covariance matrix, the computational efficiency will be greatly increased.

Given an observation sequence $O = [o_1, o_2, ..., o_t, ..., o_T]$ in duration T, the probability for state j to generate the vector o_t is computed as

$$b_j(o_t) = \sum_{k=1}^{M} w_k N(o_t; \mu_{jk}, \Sigma_{jk})$$
$$N(o_t; \mu_{jk}, \Sigma_{jk}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_{jk}|}} \exp[-\frac{1}{2}(o_t - \mu_{jk})^T \Sigma_{jk}^{-1}(o_t - \mu_{jk})] \quad \text{eq. (2-3)}$$

where w_k is the weight for mixture component k and $N(o_t;\mu_{jk}, \Sigma_{jk})$ is the multivariate Gaussian distribution of dimension n that make up the output distribution of j.

The parameters of HMM, including means, covariance matrices, state transition probabilities and initial state distributions probability, can be estimated by the Baum-Welch forward-backward training algorithm [4].

2.2.2. Basic Phonology of Cantonese

Cantonese is a monosyllabic and tonal language. Each character is pronounced as a syllable. There are about 1,800 tonal syllables. Disregarding the tone identifies, these syllables are called base syllables. The total number of base syllables is 665. A base syllable is combined with two components, an Initial and a Final. There are 19 Initials and 53 Finals in Cantonese [5]. These two basic components are called base-IF. The LSHK phonemic transcription scheme is used in our work.

Table 2-1 depicts the hierarchy of Chinese syllable. The number in the bracket is the total number of this unit. Initial onset and Final coda are optional and may not be found

in the unit. Moreover, only consonants are found in the Initial onset while vowels are contained in the nucleus.

Base Syllable (665)		
Initial (19)	Final (53)	
[Onset] (19)	Nucleus (8)	[Coda] (8)

Table 2-1: Phonologies hierarchy of Cantonese Syllable

2.2.3. Acoustic Modeling for Cantonese

The number of word is huge and thus it is difficult to use a word as the basic unit of acoustic model. Subword units are used instead and each word is mapped to a combination of subword units. In Cantonese, the subword unit is usually base-IF.

In order to better handle the co-articulation effect, context-dependent base-IF is used, which is a base-IF with a specific right context. It is called as biphone. Except silence and short pause models, all HMMs are biphone models. Here is an example of the context-independent models and biphone models corresponding to a word sequence <Sil 兩手和黃 Sil>

Context-independent models	Sil I_l F_oeng I_s F_au I_w F_o I_w F_ong
Biphone models	Sil I_l+F_oeng F_oeng+I_s I_s+F_au F_au+I_w I_w+F_o F_o+I_w I_w+F_ong F_ong+Sil Sil

Table 2-2: Context-independent and biphone models sequence for the word sequence < Sil 兩手和黃 Sil>

The prefix "I_" denotes an Initial and "F_" denotes a Final. For example, in a biphone model "I_l+F_oeng", "F_oeng" is the right context of the base phone "I_l".

The number of context-dependent models may be very large if all possible contexts are modeled. It would cause a heavy computation afford in training and recognition processes. In order to reduce the number of parameters, decision-tree based state clustering is applied so that states with similar features would be tied together [5].

2.2.4. Language Modeling

A language model gives constraints in concatenating the words to form grammatically correct strings. The probability of each pair of word is estimated by statistical approach. N-gram language model provides different probabilities to a word depending on its N-1 previous words. If N increases, the combinations of word sequence would increase exponentially. In our work, bigram language model is used. Therefore the probability that a word concatenated with its pervious word is determined.

In order to reduce the number of parameters in training and recognition process, a class-based language model has been proposed [6]. All words are first clustered into different classes. The clustering is usually based on the part of speech and also the application of system. Then, instead of finding the probability that two words are concatenated, the probability of the combination of two classes that the words belong to is used. Since the number of classes is usually much smaller than that of words, the number of parameters decreases.

References

[1] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp.254-272, 1981.

[2] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., vol. 55, no.6, pp. 1304-1312, 1974.

[3] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," J. Acoust. Soc. Amer., vol. 87, no. 4, 1990.

[4] L. R. Rahiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in proceedings of the IEEE, vol. 77, no. 2, February 1989.

[5] Y. W. Wong, "Large Vocabulary Continuous Speech Recognition for Cantonese," Mphil Thesis, The Chinese University of Hong Kong, 2000.

[6] P. F. Brown, V. J. Della Pieta, P. V. deSouza, J. C. Lai and R. L. Mercer, "Class-based N-gram Models of Natural Language," Computational Lingustics, 18(4): 467-480, 1992.

Chapter 3

Unsupervised Model Adaptation

A hidden Markov model (HMM) based speech recognition system tends to perform badly when operating in a mismatched condition. This mismatch refers to the difference in speaker's characteristics, speaking styles, accents, transmission channels or environments between training utterances and the utterance to be recognized. In order to reduce this mismatch, model adaptation is used.

In this chapter, we will focus on unsupervised model adaptation techniques. Two commonly used methods, Maximum a Posteriori (MAP) [1] and Maximum Likelihood Linear Regression (MLLR) [2], are described. Moreover, we would evaluate these techniques in three different recognition tasks and try to find the optimal parameter setting for each task.

3.1. A General Review of Model Adaptation

A speech model is usually trained for a particular condition. Such condition-specific model has better recognition performance than a general model, e.g. a speaker

independent (SI) model. Condition-specific model captures the characteristics of a speaker or acoustic condition while a general model tends to cover many different speakers or conditions simultaneously. Large amount of acoustic data is required to train a condition-specific model. This is not practical in most applications. Therefore, model adaptation is needed to make a general model become a condition-specific model.

Model adaptation attempts to adjust the model parameters with a small amount of data so that the adapted model better suits the new operating condition and the mismatch between model and data can be reduced.

Speaker variation is one of the major types of condition mismatch being encountered in speech recognition. Model adaptation can minimize such mismatch and it is one of the approaches to achieve speaker adaptation. Another approach is speaker normalization [4]. Speaker normalization attempts to transform the short-time features of the new speaker's speech in the feature extraction part. The input speech is normalized so that the mismatch of feature between training speakers and the new speaker is minimized. However, it is difficult to model the characteristic of human vocal tract which limits the development of this technique.

Our research deals with HMM-based speech recognition system. There are two ways to perform model adaptation: supervised adaptation and unsupervised adaptation.

3.1.1. Supervised and Unsupervised Adaptation

If the true content of the adaptation data is known, supervised adaptation is done as shown in Figure 3-1.

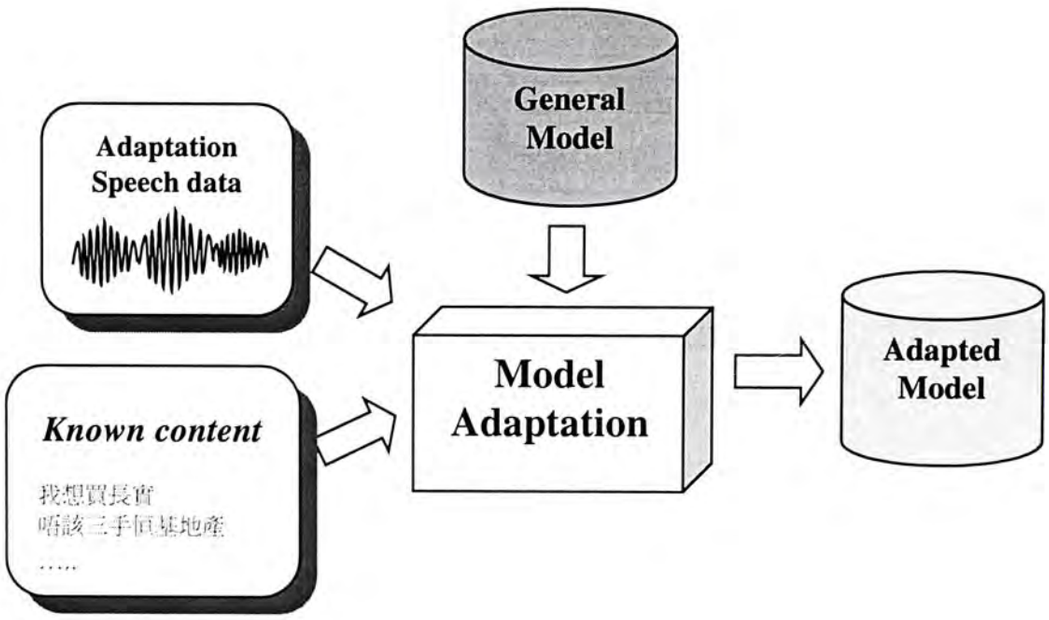


Figure 3-1: The flowchart of supervised adaptation

If the content is not provided, the adaptation must be done in an unsupervised manner. In this case, the adaptation data firstly pass through a speech recognition system so as to obtain a hypothesis of the content. The general model is then adjusted based on the hypothesis. The flowchart is shown in Figure 3-2.

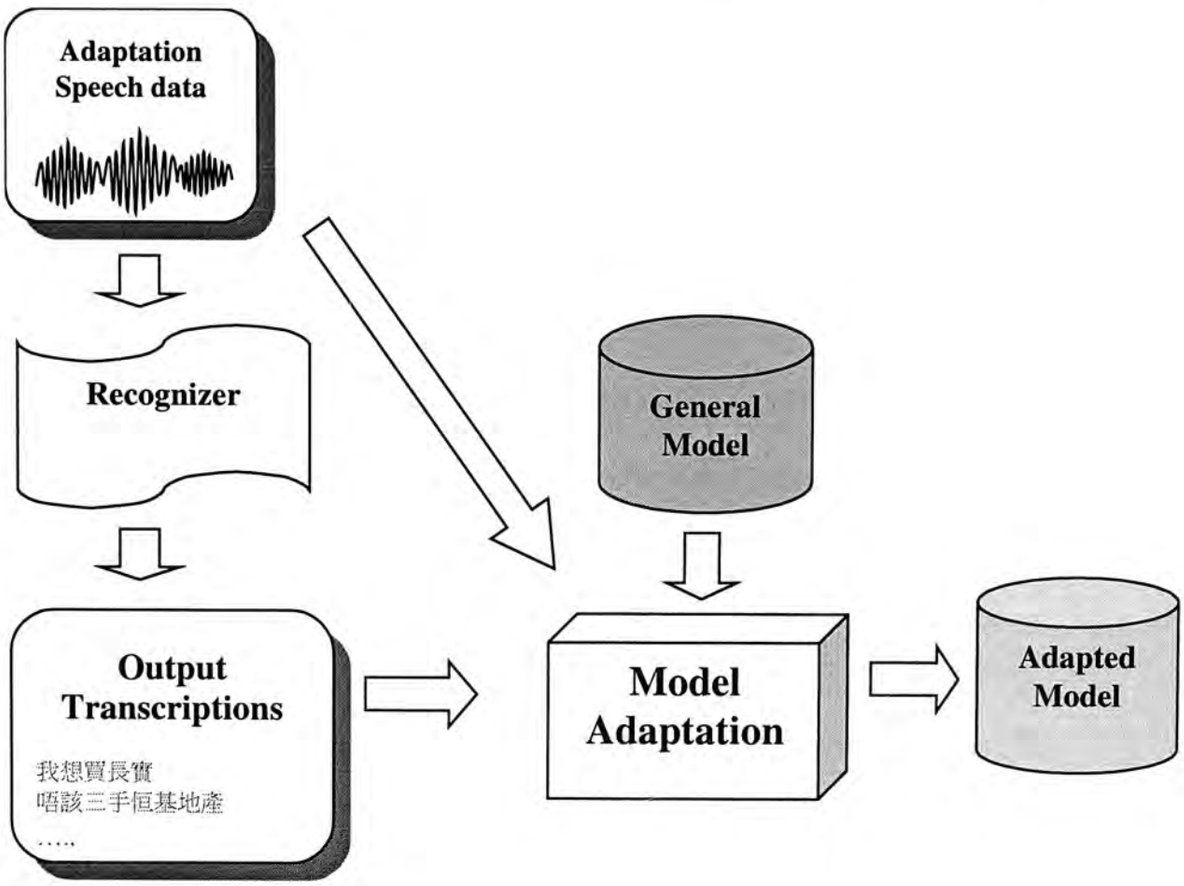


Figure 3-2: The flowchart of unsupervised adaptation

For unsupervised adaptation, the effectiveness of adaptation depends on the accuracy of the recognizer. Recognition errors mean the incorrect mapping between observation vectors and the models to be adapted. The model would be adjusted by the unrelated observation vector. This error would affect the estimation of the adapted model. Therefore, unsupervised adaptation usually performs worse than the supervised adaptation does.

Unsupervised adaptation is usually used since the content of the input speech is unknown in most applications. It would be better if there is an automatic mechanism to filter out the recognition errors. Confidence measure is one of such techniques. The concept of confidence measure is related to the correctness and reliability of a

recognition hypothesis and hence facilitates certain kind of supervision for unsupervised adaptation. In this way, the quality of adaptation data could be better controlled [3]. This will be discussed in the next chapter.

Another practical problem in adaptation is the insufficient amount of adaptation data. It is difficult to get a large amount of data for adaptation in most applications. Therefore, a good adaptation technique should be able to extract as much information as possible from limited adaptation data. One possible way is to make use of the N-best recognition outputs.

3.1.2. N-Best Adaptation

Nowadays, many recognizers can generate not only the most possible hypothesis but also N hypotheses ranked by their path scores. These N hypotheses both carry the information about the input speech that is valuable to the adaptation. Therefore, it is advantageous to involve not only the first best hypothesis but also some lower-rank hypotheses. It is called N-Best Adaptation.

To ensure the balanced contributions from different utterances, each hypothesis would be weighted by a certain factor, which is equal to the reciprocal of the number of hypotheses. Therefore, the sum of all weightings is 1. In this way, the contributions of an utterance with only a single hypothesis and one with multi-hypotheses are the same.

After a general review on the concept of model adaptation, two major model adaptation techniques, Maximum a Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR), will be discussed in details.

3.2. MAP

MAP is based on the Bayesian estimation with prior knowledge of model incorporated in the adaptation. Since a posterior probabilities are maximized in this approach, it is called Maximum A Posterior (MAP). Being different from Maximum Likelihood (ML) estimation, Bayesian estimation assumes that the parameters are random vectors that carry prior information. The adaptation data are used to adjust the values of the parameters.

In parameter estimation, we need to find out the optimal parameter that maximizes a posterior probability given observation O. Using the Bayes Theorem, the a posterior probability is as

$$p(\lambda \mid O) = \frac{p(O \mid \lambda) g(\lambda)}{p(O)} \qquad \text{eq. (3-1)}$$

where λ denotes the parameter of the system and the $g(\lambda)$ is a priori distribution of the parameter.

Then the MAP estimate λ_{MAP} can be defined as follow.

$$\lambda_{MAP} = \arg\max_{\lambda} p(x \mid \lambda) g(\lambda) \qquad \text{eq. (3-2)}$$

The MAP estimates of the Gaussian mixture components can be found by applying the EM algorithm. Adapted mean is computed as follow.

$$m_k = \frac{\tau_k \mu_k + \sum_{t=0}^{T} c_{kt} x_t}{\tau_k + \sum_{t=0}^{T} c_{kt}} \qquad \text{eq. (3-3)}$$

where $c_{kt} = \frac{\omega_k N(x_t \mid \lambda_k)}{\sum_{l=1}^{K} \omega_l N(x_t \mid \lambda_l)}$

In this equation, μ_k is prior mean which is usually defined as the general mean. τ_k is the weight factor, controlling the balance between the initial model and adaptation data. Usually the weight factor is the same among all models.

Graphically, the relation among the observation vector x_t, the general mean vector μ_k and the adapted mean m_k can be illustrated below.

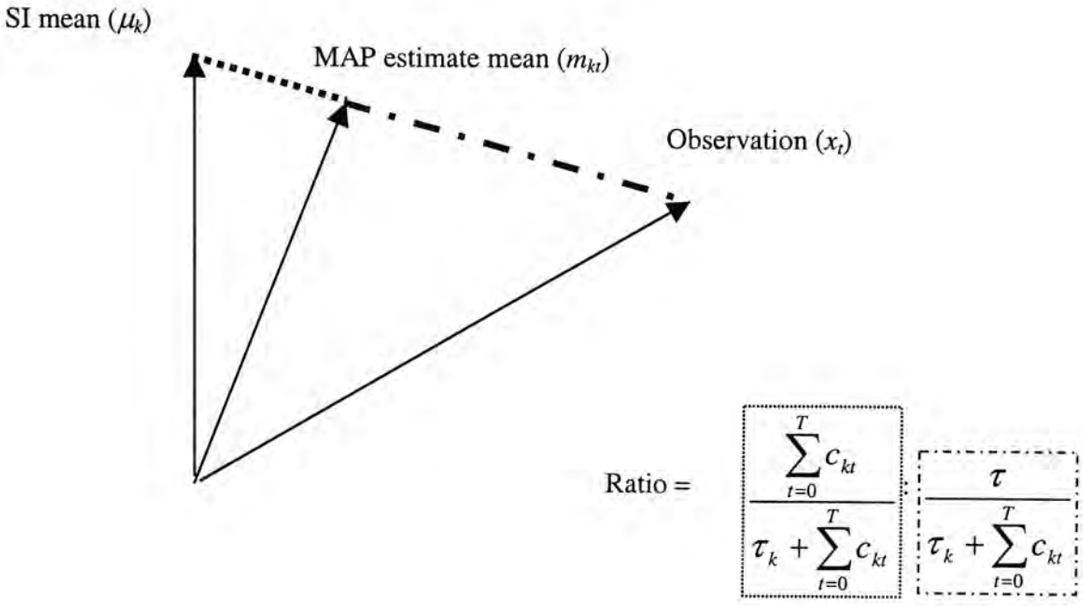


Figure 3-3: The geometric description of MAP

The MAP adapted mean can be simplily viewed as the linear combination of the general-condition mean and the observation vector. By increasing the amount of adaptation data, $\sum_{t=0}^{T} c_{kt} x_t$ is comparable or even much larger than the weight factor τ_k. Then the effect of the observation vector would be dominant and the MAP estimation is very similar to the ML estimation. On the other hand, the adapted mean remains close to the general mean when the adaptation data are insufficient.

The disadvantage of MAP is that it involves only the parameters of the models that are observed in the adaptation data. For a system with large number of models, the required adaptation data are fairly large in general. In order to tackle this problem, some reserachers proposed an additional training. It is believed that those model vectors close in space would have similar degree of adjustment. According to the movement of the surrounding adapted models, the unadapted models are also updated [5].

3.3. MLLR

Maximum Likelihood Linear Regression (MLLR) uses linear transformation to adapt HMM parameters based on the maximum likelihood criterion [2]. Each transformation is applied to a group of HMM parameters. It is firstly applied in the transformation of mean vectors and further extended to variances [6]. By sharing transformations and data, this method can lead to performance improvements with relatively small amounts of adaptation data.

3.3.1. Adaptation Approach

The adaptation of a mean vector is done by applying a transformation matrix W_s to the extended mean vector $\xi_s = [1, \mu_1, \ldots, \mu_N]$, i.e.

$$\mu_s' = W_s \xi_s \qquad \text{eq. (3-4)}$$

Therefore, mean vector in the probability density function $b_j(o_t)$ in eq. (2-3) is replaced by the product of transformation and mean vector. The function can be expressed as follow.

$$b_s(o) = \frac{1}{(2\pi)^{\pi/2} |C_s|^{1/2}} e^{-1/2(o-W_s\xi_s)' C_s (o-W_s\xi_s)} \qquad \text{eq. (3-5)}$$

If each Gaussian distribution has its own transformation matrix, essentially a complete re-estimation of the means has been done. However, the amount of data is also same as that required for the training process. When adaptation data is only in small amount, some distributions may not have any data or have small amount of data but not enough for estimating a good transformation. Therefore, grouping distributions and thus sharing transformation is a good solution. For an unseen distribution, it can still be adapted by the shared transformation.

Gaussian distributions are tied into several groups first and the transformation is estimated using data from all distributions in the same group. It is crucial to determine the number of groups dynamically according to the amount of adaptation data. This is done based on a regression class, which will be discussed in Section 3.4.4.

3.3.2. Estimation of MLLR regression matrices

The total likelihood of HMM sequence λ generating the observation sequence O is

$$F(O \mid \lambda) = \sum_{\theta \in \Theta} F(O, \theta \mid \lambda) \qquad \text{eq. (3-6)}$$

where Θ is the set of all possible state sequences of length T and $F(O, \theta \mid \lambda)$ is the likelihood of O with the state sequence θ given the model λ.

To maximize $F(O \mid \lambda)$, an auxiliary function $Q(\lambda \mid \lambda')$ is defined as,

$$Q(\lambda, \lambda') = \sum_{\theta \in \Theta} F(O, \theta \mid \lambda) \log(F(O, \theta \mid \lambda') \qquad \text{eq. (3-7)}$$

Maximizing $Q(\lambda \mid \lambda')$ over λ improves λ' in the sense of increasing the likelihood $P(O \mid \lambda)$. Our ultimate goal is to estimate W_s which is only included in the function $b_j(o_t)$. Therefore, the auxiliary function can be rewritten as

$$Q(\lambda, \lambda') = \text{constant} + \sum_{\theta \in \Theta} \sum_{t=1}^{T} F(O, \theta \mid \lambda) \log(b_{\theta t}(o_t)) \qquad \text{eq. (3-8)}$$

Defining S as the set of all state distributions in the system and expanding $\log(b_{\theta t}(o_t))$, the auxiliary function becomes

$$Q(\lambda, \lambda') = \text{constant} + \frac{1}{2} F(O \mid \lambda) \sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_s(t) \left[n \log(2\pi) + \log \mid C_j \mid + (o_t - W_j \xi_j)' C_j^{-1} (o_t - W_j \xi_j) \right] \qquad \text{eq. (3-9)}$$

where $\gamma_s(t) = \dfrac{1}{F(O \mid \lambda)} \sum_{j=1}^{S} F(O, \theta_j \mid \lambda)$

$\gamma_s(t)$ is the a posteriori probability of occupying state s at time t given the observation sequence.

Differentiating $Q(\lambda|\lambda')$ with respect to W_s and equates to zero to find the maximum of $Q(\lambda|\lambda')$.

$$\frac{d}{dW_s}Q(\lambda,\lambda') = F(O \mid \lambda)\sum_{t=1}^{T}\gamma_t(t)C_s^{-1}(o_t - W_s\xi_s)\xi_s' = 0 \qquad \text{eq. (3-10)}$$

$$\sum_{t=1}^{T}\gamma_s(t)C_s^{-1}o_t\xi_s' = \sum_{t=1}^{T}\gamma_s(t)C_s^{-1}W_s\xi_s\xi_s' \qquad \text{eq. (3-11)}$$

Since those transformation matrices are shared by R states $\{s_1, s_2, \ldots s_R\}$, the summation will be performed over all R states. The equation becomes

$$\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{sr}(t)C_{sr}^{-1}o_t\xi_{sr}' = \sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{sr}(t)C_{sr}^{-1}W_s\xi_{sr}\xi_{sr}' \qquad \text{eq. (3-12)}$$

To derive a re-estimation formula for the tied case, the above equation is rewritten as

$$\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{sr}(t)C_{sr}^{-1}o_t\xi_{sr}' = \sum_{r=1}^{R}V^{(r)}W_sD^{(r)} \qquad \text{eq. (3-13)}$$

$$V^{(r)} = \sum_{t=1}^{T}\gamma_{sr}(t)C_{sr}^{-1}$$
$$D^{(r)} = \xi_{sr}\xi_{sr}'$$

If the left-hand side is denoted by the n×(n+1) matrix Y with elements y, the individual matrix elements of $V^{(r)}$, W_s and $D^{(r)}$ are denoted by $v_{ij}^{(r)}$, w_{ij} and $d_{ij}^{(r)}$ respectively.

$$y_{ij} = \sum_{p=1}^{n}\sum_{q=1}^{n+1} w_{ij}[\sum_{r=1}^{R} v_{ip}^{(r)} d_{qj}^{(r)}] \qquad \text{eq. (3-14)}$$

Since a diagonal covariance matrix is used, then

$$\sum_{r=1}^{R} v_{ip}^{(r)} d_{qj}^{(r)} = \begin{cases} \sum_{r=1}^{R} v_{ip}^{(r)} d_{qj}^{(r)} & \text{when i = p} \\ 0 & \text{when i} \neq \text{p} \end{cases} \qquad \text{eq. (3-15)}$$

Defining a matrix $G^{(i)}$ has (n+1)×(n+1) elements $g_{jq}^{(i)}$

$$g_{jq}^{(i)} = \sum_{r=1}^{R} v_{ii}^{(r)} d_{qj}^{(r)} \qquad \text{eq. (3-16)}$$

Therefore, eq. (3-13) can be simplified

$$y_{ij} = \sum_{p=1}^{n} w_{iq} g_{jq}^{(i)} \qquad \text{eq. (3-17)}$$

Then

$$w_i = G^{(i)^{-1}} y_i \qquad \text{eq. (3-18)}$$

where w_i and z_i are the i^{th} rows of W_s and Z_s respectively. For N component mean, we need to find the inverse of N matrices $G^{(i)}$ for each transformation matrix W_s.

3.3.3. Least Mean Squares Regression

Least Mean Square (LMS) Regression is a special case of MLLR. It assumes that distributions have identical covariances when they share same transformation. Thus, the eq. (3-10) can be further simplified to

$$\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{sr}(t)o_t\xi_{sr}' = \sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{sr}(t)W_s\xi_{sr}\xi_{sr}' \qquad \text{eq. (3-19)}$$

If the Viterbi algorithm is used for search, $\gamma_s(t)$ acts as the selection of adaptation data.

$$\gamma_{sr}(t) = \begin{cases} 1 & o_t \text{ assigned to state distributions } s_r \\ 0 & \text{otherwise} \end{cases} \qquad \text{eq. (3-20)}$$

Let the left-hand side of eq. (3-19) is denoted by n×(n+1) Z. The estimation of transformation can be written as

$$W_s = H^{-1}Z \qquad \text{eq. (3-21)}$$

where $H = \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{sr}(t)\xi_{sr}\xi_{sr}'$

Unlike the standard MLLR, only the inverse of a matrix H is required in the estimation of W_s. Thus the computation effort can be much reduced. However, the information of covariance is neglected. In [2], this information is found to be useful for adaptation. We try to evaluate the contribution of the information of covariance by comparing these two methods in experimental result section.

3.3.4. Number of Transformations

Clustering of distributions and amount of data are the major concerns when deciding the number of transformations. The regression class can give a satisfactory performance. It groups the data based on their spectral distances and decides the number of transformations accounting to the amount of adaptation data. If there are

only small amount of data, a global adaptation transformation can be generated. It will be applied to every Gaussian mixture in the model set. With the increase in amount of data, the number of transformations can be increased. Then, each transformation becomes more specific to a group of Gaussian mixtures.

The regression class tree is used to categorize Gaussian distributions into groups. The tying of class is according to the amount of data. If the amount of adaptation data in a class were insufficient, it would be tied to another class and share the transformation. The approach for defining regression class is usually based on the clustering of Gaussian mixtures. The mixtures that are close in acoustic space will be grouped together. The tree is constructed with a centroid splitting algorithm.

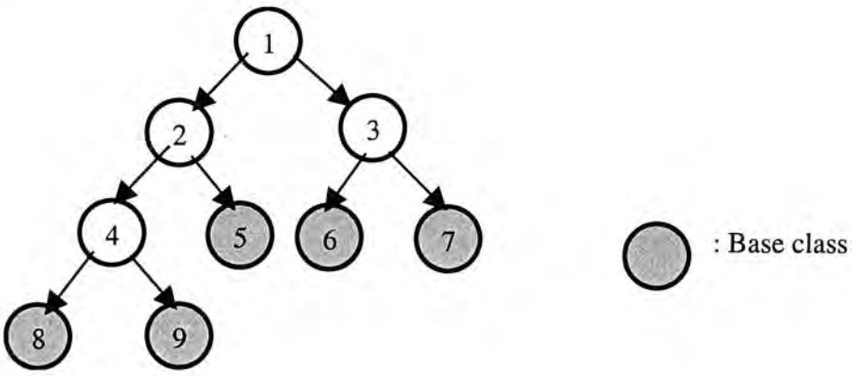


Figure 3-4: A binary regression class tree

The terminal nodes are called base classes which define the final grouping. Therefore, each Gaussian mixture belongs to a specific base class. Figure 3-4 shows a regression class tree. The gray circles are the base classes of the regression class tree.

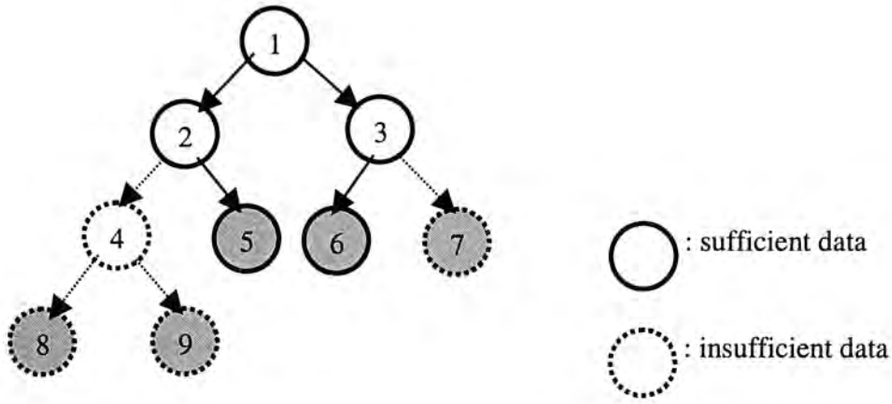


Figure 3-5: An example of a binary regression class tree

An example is shown in Figure 3-5. In nodes 1, 2, 3, 5 and 6, there are sufficient adaptation data for generation of their own transformation matrices. However, for those nodes that do not have enough data, they would be tied with other nodes. For example, node 7 doesn't have enough data. It will look for its parent node, node 3, which has pooled the data into node 6 and 7. Since the data in node 3 is enough to form a transformation matrix, this transformation matrix would be shared by node 6. It is the same case for nodes 4, 8 and 9. The base node with insufficient data would look up the tree until it can find a transformation matrix to share with.

3.4. Experiment Results

We will conduct several experiments in three different tasks. There are three objectives. I) Comparison between standard MLLR and LMS MLLR, II) Finding the optimal parameter setting of MLLR, III) Comparison between MLLR and MAP. Experiments are carried out on the following tasks.

- **Task 1: Domain specific microphone speech recognition**

It is a domain-specific recognition system which is designed for stock information inquiry. Since the vocabulary used in stock market is much less than that in newspaper or general domain, the lexicon is much smaller. It is for the microphone speech input. The environment is quiet and the channel noise in microphone is relatively small.

In the microphone recognition system, the acoustic models are trained using CUSENT, which is a large microphone speech corpus developed by DSP lab. The total amount of data is about 20 hours and the training speakers are over 68. The acoustic models are 997 decision tree based cross-word biphones. The number of Gaussian mixtures at each state is 16.

The lexicon contains only 1,125 words. The language model is trained by 2095 queries.

The testing set contains 500 sentences recorded by 5 speakers. First 20 sentences of each speaker are used for adaptation and the rest 80 sentences are for evaluation.

- **Task 2: Domain-specific telephone speech recognition**

Similar to Task 1, it is designed for stock information inquiry. Therefore, the language model is same as Task 1. The acoustic model is trained by telephone speech. It is used for telephone speech recognition. Since the telephone speech

would be noisier than microphone speech due to the channel noise and environment noise, we can evaluate whether the model adaptation can perform well in noisy but domain-specific condition.

In the telephone recognition system, the acoustic models are trained using CUCall which is a large telephone speech corpus developed by DSP lab in CUHK. The total amount of data is about 80 hours and the training speakers are over 1000 peoples. The acoustic models are 956 decision tree based cross-word biphones. The number of Gaussian mixtures at each state is 16.

The testing set contains 600 sentences recorded by 6 speakers through telephone line. First 20 sentences of each speaker are used for adaptation and the rest 80 sentences are for evaluation.

- **Task 3: Telephone recognition system for general domain LVCSR**

Unlike the domain-specific recognition system, the vocabulary is much larger in this task. Its language models are trained by the newspaper data. It is much more complicated than domain-specific language model and the recognition accuracy is much lower. The acoustic model is trained by telephone speech. Therefore, we will evaluate whether the model adaptation can perform well and the confidence measure can correctly determine the reliability of adaptation data in low recognition condition.

The acoustic model is same as Task 2.

The lexicon is much larger than that of stock information inquiry system. It contains 6,449 words. The language model is trained by 2095 queries.

The testing set contains 275 sentences. Each speaker records 55 sentences. First 20 sentences of each speaker are used for adaptation and the rest 35 sentences are for evaluation.

In Task 1 and 2, the recognition systems are designed for same domain but different speaking environments. One is for using in telephone channel and another is for using in microphone channel. The speaking environment is the main difference between these two tasks.

In Task 2 and 3, the recognition systems are both for the telephone speech but they are using in different domains.

We would like to evaluate our method in different conditions in order to find out the limitations or pre-requirements.

3.4.1. Standard MLLR versus LMS MLLR

A good adaptation technique is expected to have a good balance between accuracy and computation required. A comparison between these two methods is carried out in [2] also. The results shown that standard MLLR requires larger effort of computation but gives better performance than LMS MLLR does. Moreover, the word error rate had been reduced from 4.3% to 2.8% after applying standard MLLR when the adaptation data of 40 sentences and 15 transformations are used, which is 25% better than that using LMS MLLR. However, we need to calculate 38 matrix inverses and 39×(number of frame) floating point multiplications more when using the standard MLLR if 39-dimension feature vector is used.

In order to choose a suitable adaptation technique which has a good balance, we will evaluate these two methods in Task 1 and 2. The experiment will be conducted with 1-best and 10-best adaptations using different numbers of transformation. The word error rate (WER) and real time factors (RTF) will be found. RTF is the ratio of the time for the recognition process to the length of the input utterance.

Task 1

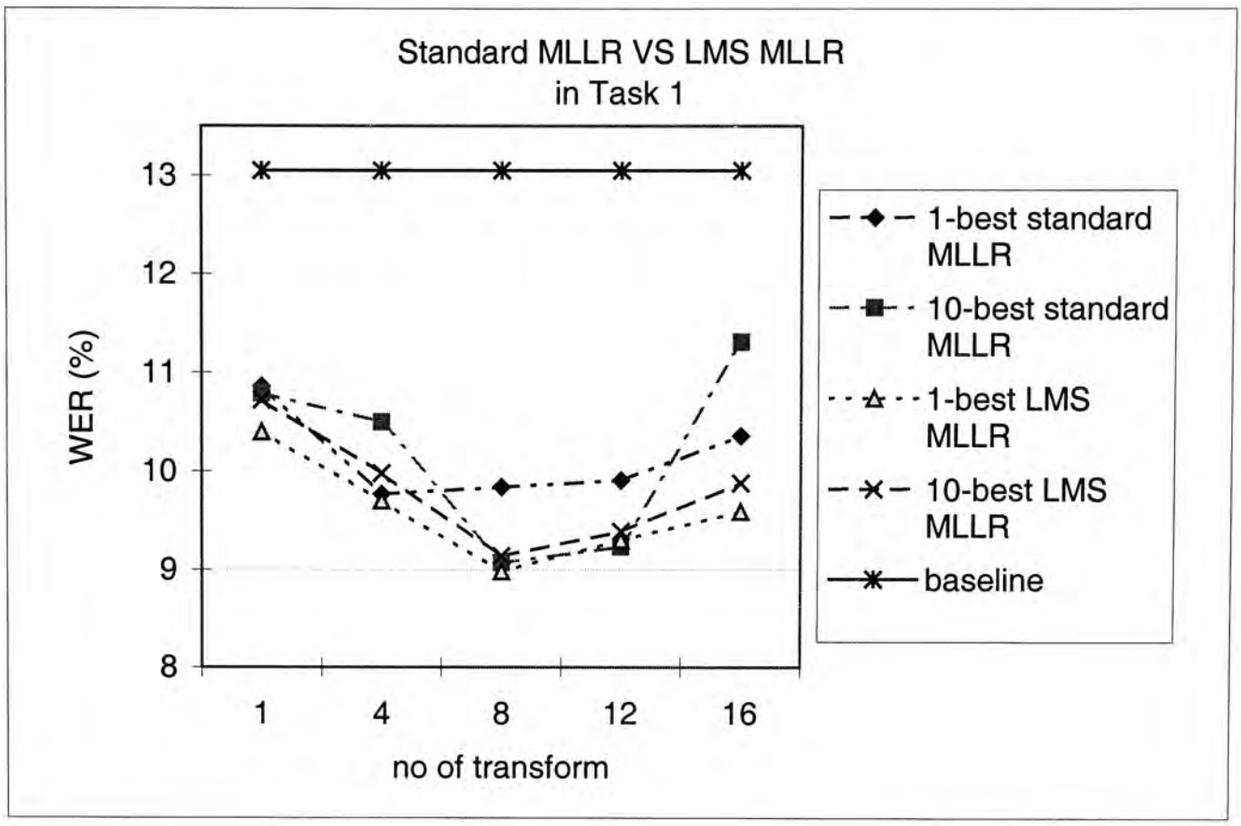


Figure 3-6: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformation in Task 1.

	Standard MLLR	**LMS MLLR**
1-best	3.63	0.93
10-best	5.71	1.73

Table 3-1: The RTF per sentence of different technique in Task 1.

Task 2

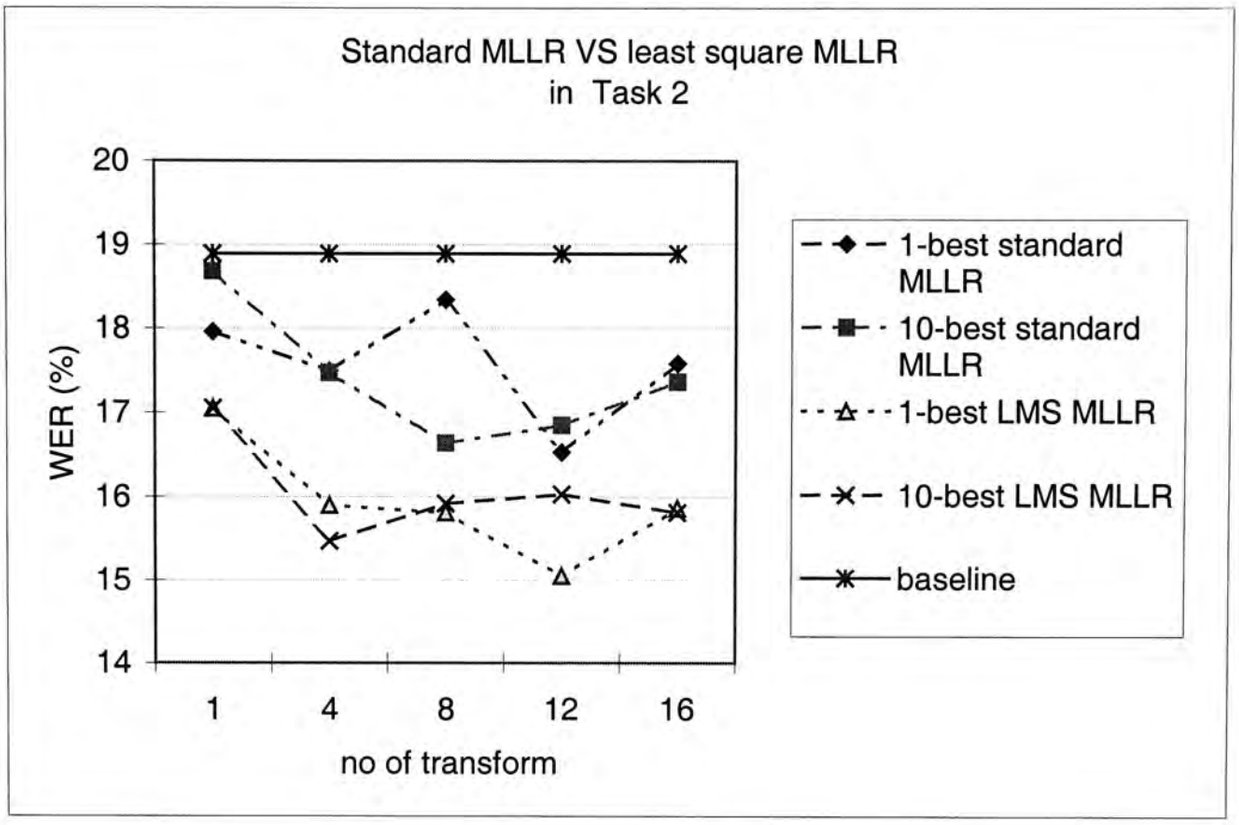


Figure 3-7: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformation in Task 2.

	Standard MLLR	**LMS MLLR**
1-best	2.37	0.96
10-best	4.70	1.27

Table 3-2: The RTF per sentence of different techniques in Task 2.

The LMS MLLR generally has the best performance in all cases. In Task 2 the standard MLLR performs much worse than the LMS MLLR. We can find the separation between two sets of curves is large. It shows the information of covariances may not be useful. This is quite different from the results in [2]. We would try to increase the adaptation data to 40 sentences so that it is equal to the optimal setting in [2], and see whether the expected result can be observed.

Task 1

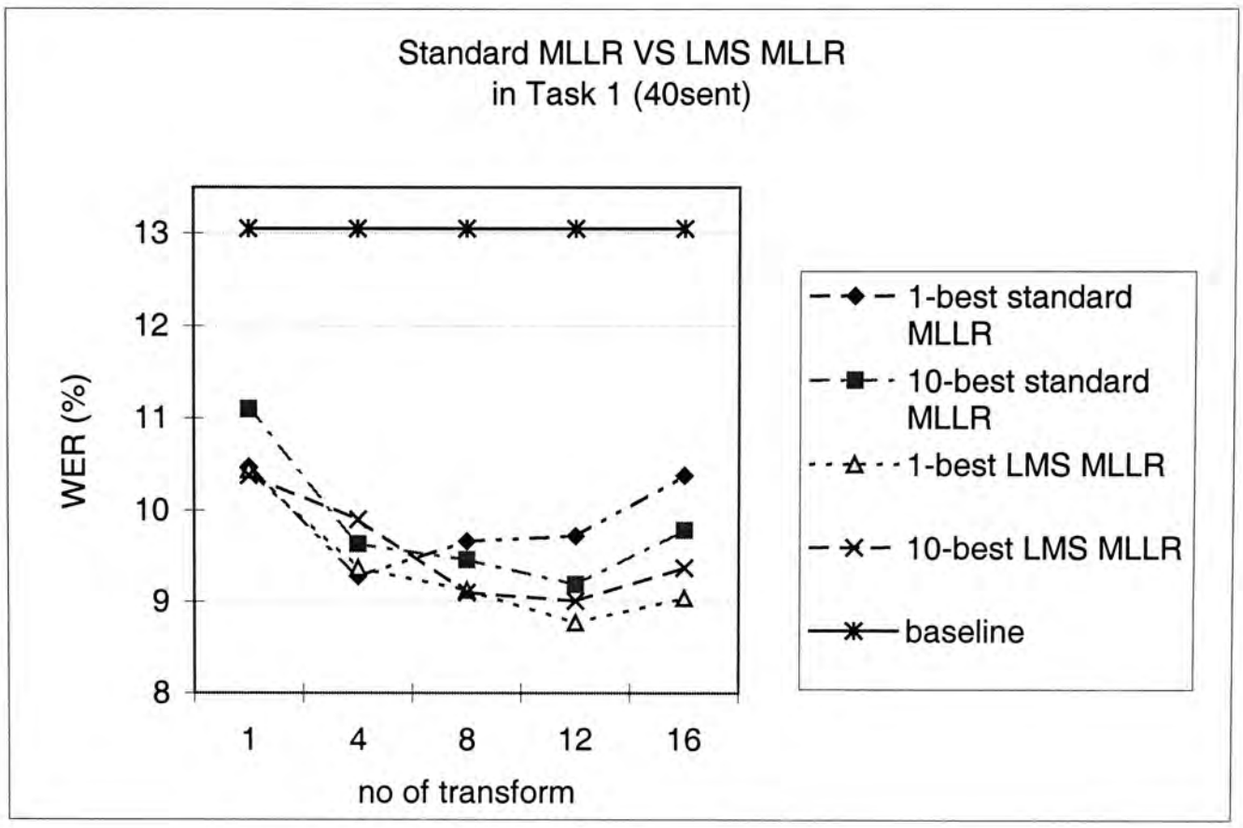


Figure 3-8: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformations in Task 1 (40 adaptation sentences).

	Standard MLLR	**LMS MLLR**
1-best	3.03	0.92
10-best	5.54	1.74

Table 3-3: The RTF of different technique in Task 1 (40 adaptation sentences).

Task 2

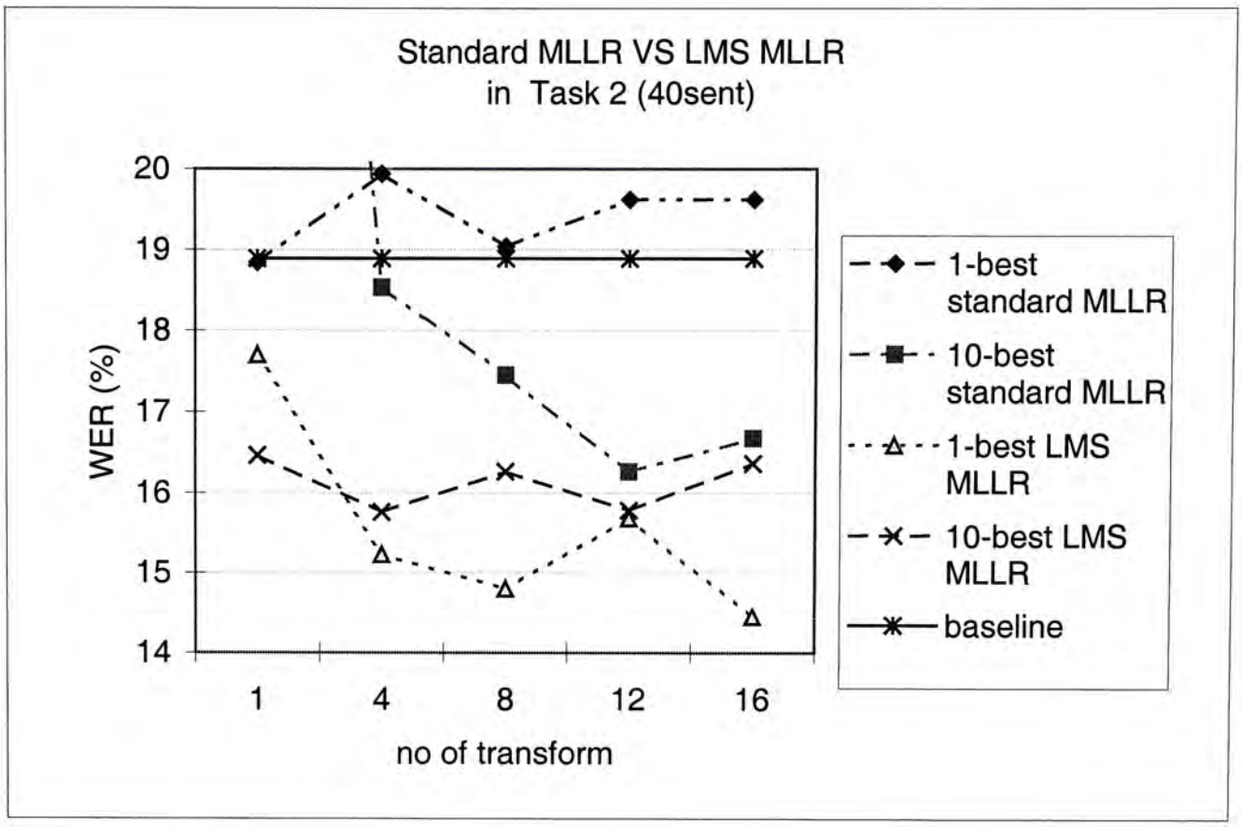


Figure 3-9: The WER (%) after applying standard MLLR and LMS MLLR with different numbers of transformation in Task 2 (40 adaptation sentences).

	Standard MLLR	**LMS MLLR**
1-best	2.28	0.89
10-best	4.83	1.38

Table 3-4: The RTF per sentence of different technique in Task 2 (40 adaptation sentences).

The results in the two tasks are quite different. In Task 1, the result is similar to the pervious one. The LMS MLLR performs slightly better than the standard MLLR. But in Task 2, the performance of 1-best standard MLLR performs even worse than the

baseline. In 10-best adaptation using the global transformation, the WER is even as high as 30% which is much larger than the baseline.

The performance of standard MLLR does not become better when the adaptation data increases, but even worse. The implementation techniques and amount of adaptation are similar to that in [2]. The major difference between two tasks is the recognition environment and adaptation data. The unexpected result may be caused by this difference.

The covariance is sensitive to noise. The standard MLLR performs badly in Task 2 where it is a telephone recognition system. In the meanwhile, the standard MLLR has a similar performance with the LMS MLLR in Task 1 where it is a microphone recognition system. The telephone speech data is much noisier than the clean speech data. Those noises would significantly degrade the usefulness of the information of the covariances. It may even induce the error in the estimation of the transformation and so the adaptation error. In paper [2], the recognizer is trained by the clean-speech data and the baseline WER is low. The information of covariance can be isolated from the noise. The inclusion of covariance becomes useful and necessary.

Other than the accuracy, the computation time is also concerned. In a 1-best adaptation, the computation time of standard MLLR is about 3 fold of that of the LMS MLLR. It is even nearly 4 fold in the 10-best adaptation. It is costly to use the standard MLLR.

Since LMS MLLR is computation-save and more robust even the acoustic model is trained in the noisy environment, it will be used as our default MLLR method in the coming experiments.

3.4.2. Effect of the Number of Transformations

More transformations may improve the adaptation. However, the transformation may be wrongly estimated if the adaptation data is not enough. Therefore, we need to find out the optimal number of transformation. The number of transformation is dependent on the quantity of adaptation data as well as the application. Therefore we try to evaluate it in three tasks. Moreover, we would compare the result using different numbers of transformation in 1-best, 10-best and 50-best unsupervised adaptation as well as supervised adaptation are.

Task 1

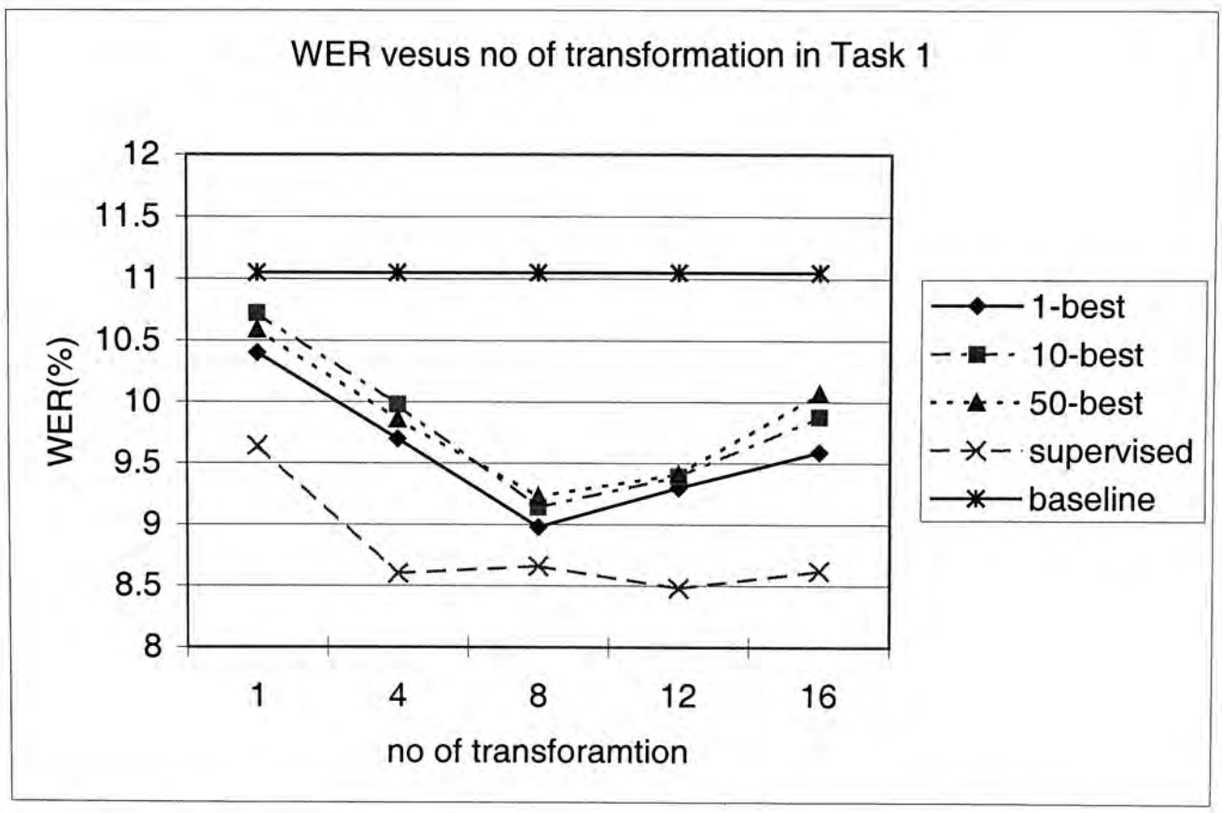


Figure 3-10: The WER (%) after applying supervised adaptation and unsupervised adaptation using N-best hypotheses in Task 1.

The optimal number of transformation is 8. The trends of the performances of 1-best, 10-best and 50-best adaptation are similar when the number of transformation increases. 1-best adaptation is slightly better than the other two.

Task 2

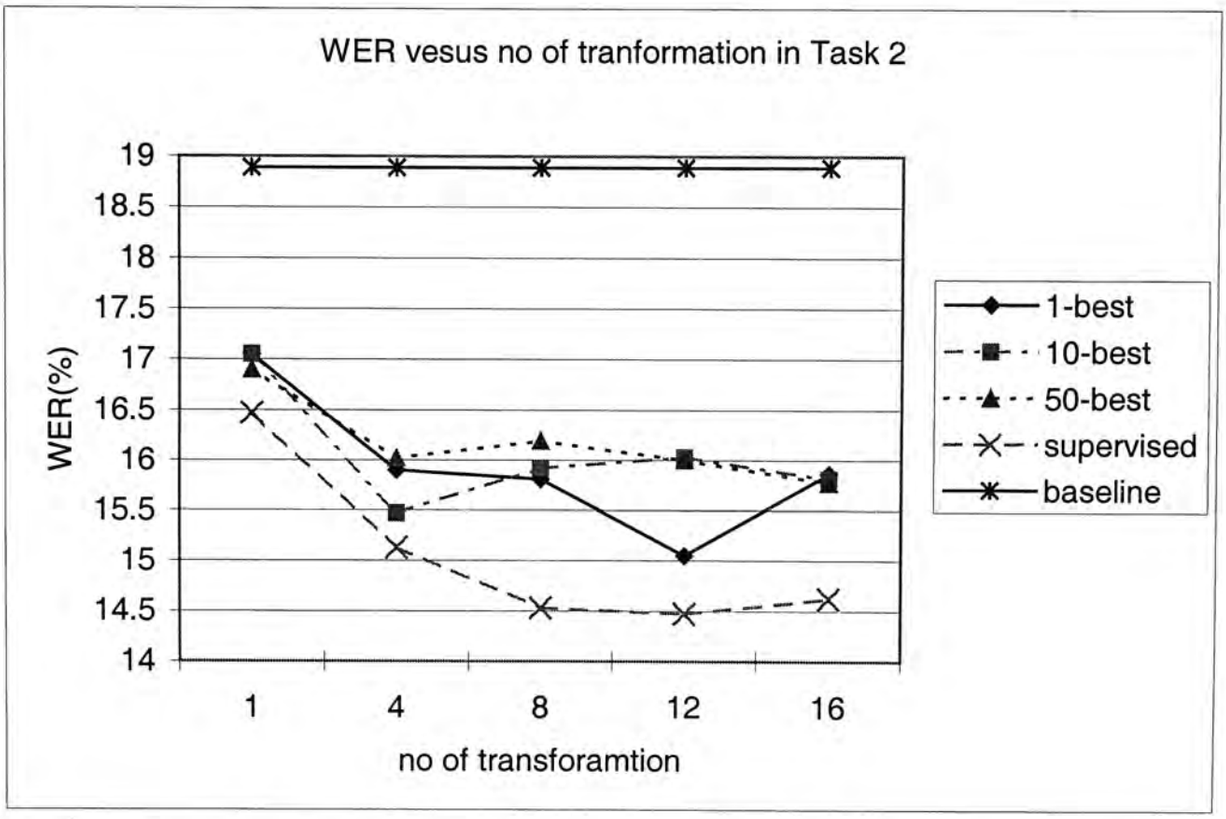


Figure 3-11: The WER (%) after applying supervised adaptation and unsupervised adaptation using N-best hypotheses in Task 2.

There is a significant improvement when we increase the number of transformation from 1 to 4. It shows that the multi-transformation can provide more specific information for particular group of distribution when the adaptation data is enough. The best performance in 1-best adaptation and supervised adaptation can be observed when 12 transformations are used. There is more than 4% improvement. However, the WER cannot be further improved in 10-best and 50-best adaptation case when

increasing the number of transformation. Since those adaptation data in the 1-best hypothesis would also be included in the 10-best and 50-best hypotheses, the problem is caused by the adaptation data in the lower-rank hypotheses. There are not only useful data but also errors in the lower-rank hypotheses. Those errors would significantly offset the benefit from the increase in the useful data. So we try to propose the confidence measure to remove those errors in order to improve the performance. We will discuss it in the coming chapter.

Even the application task is as same as that in Task 1 and so the distribution of adaptation data would be similar, the result is different. It is because the acoustic models are different in two settings. Therefore, their regression class trees are different too. It makes the distributions of the data in the tree is different.

Task 3

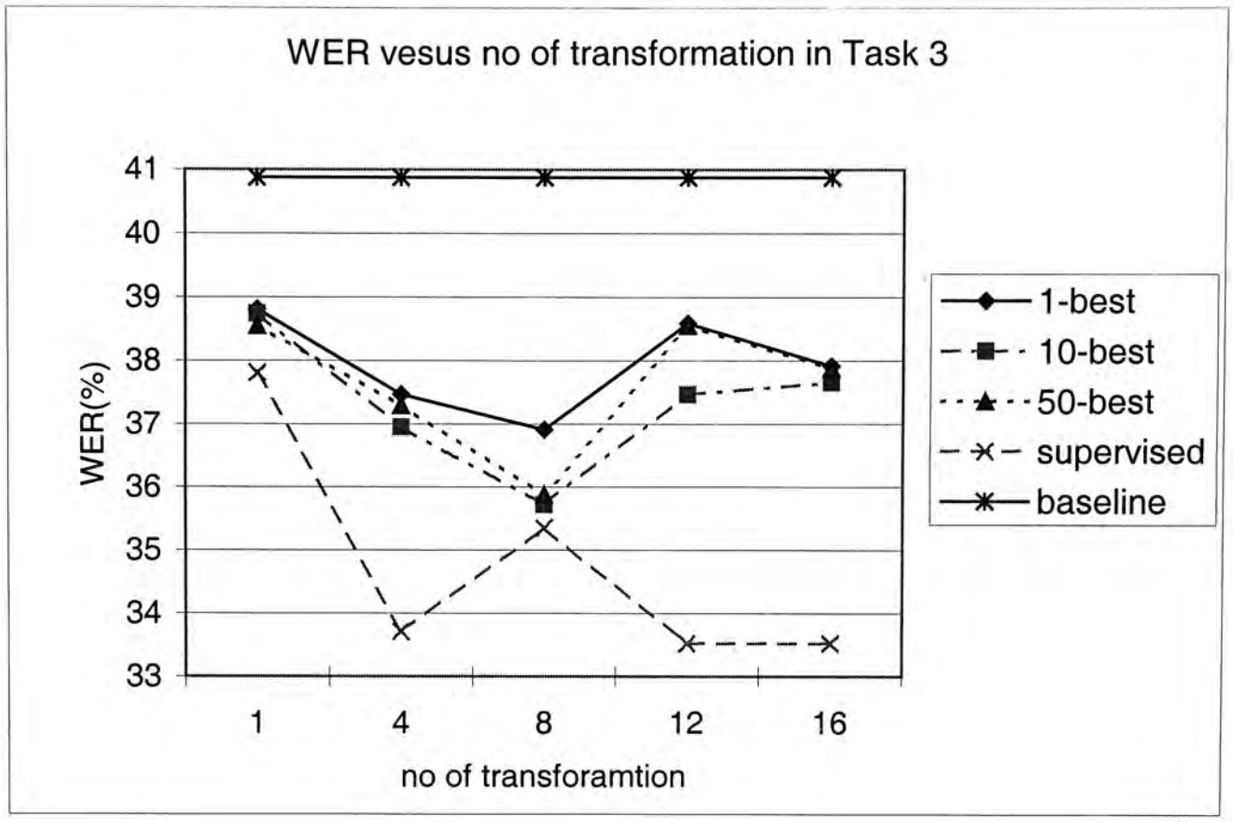


Figure 3-12: The WER (%) after applying supervised adaptation and unsupervised adaptation using N-best hypotheses in Task 3.

The optimal number of transformation is 8. The consistent trend is observed in 1-best, 10-best and 50-best adaptation.

3.4.3. MAP Vs. MLLR

In this section, we try to compare MAP and MLLR in Task 2. These two techniques have their own advantages. MLLR can adapt all parameters in model set even those parameters have not got any adaptation data, while MAP can give a more specific and precise adaptation for a large amount of adaptation data. There are some tuning factors

in these two techniques, the weight factor τ (tau) in MAP and the number of transformation in MLLR. Therefore, we try to adjust these tuning factors and compare their results.

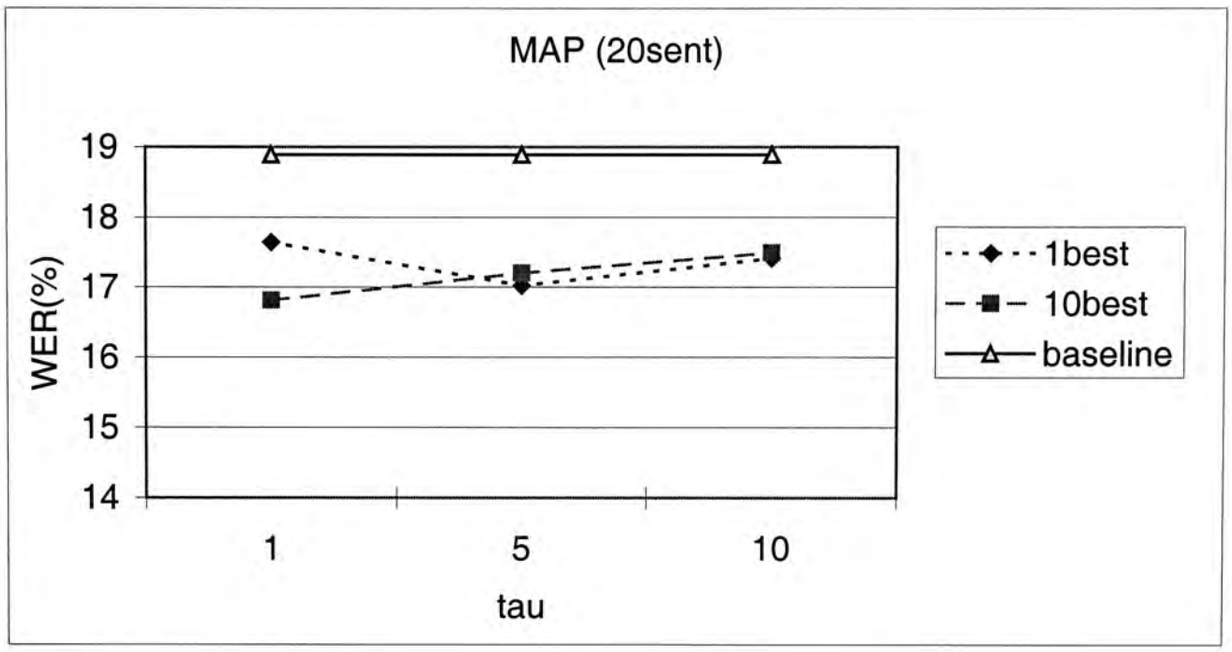


Figure 3-13: The WER(%) of MAP with 20 adaptation sentences

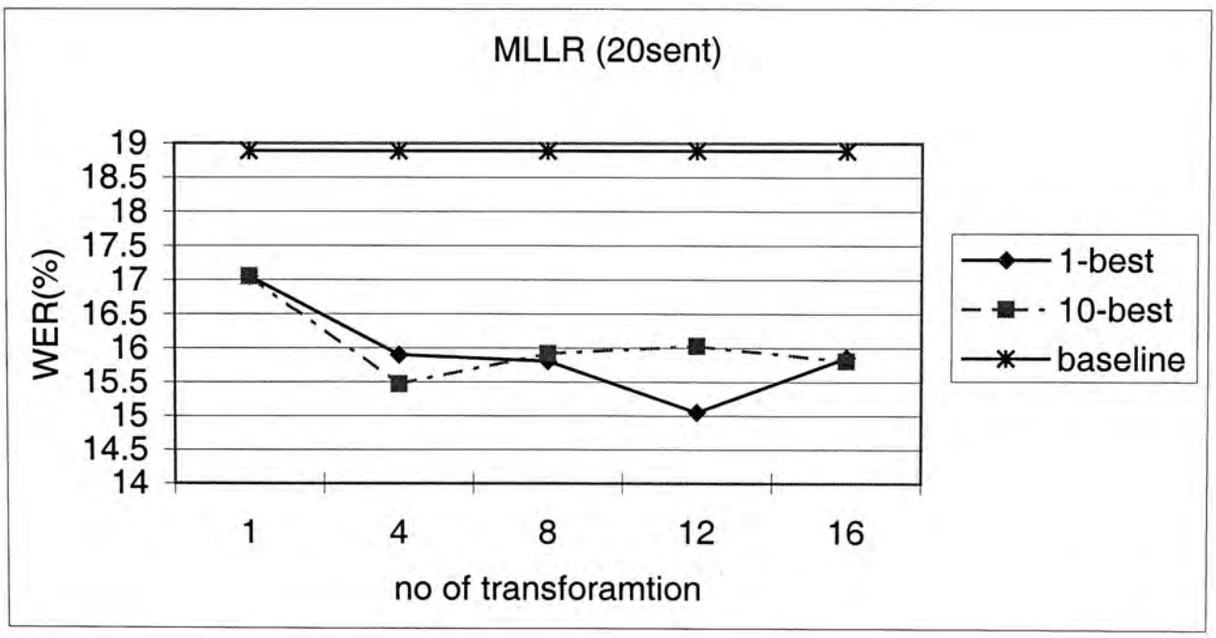


Figure 3-14: The WER(%) of MLLR with 20 adaptation sentences

Besides using the global transformation, MLLR outperforms MAP in all other cases. The reason may be that 20 sentences adaptation data may be not enough for the MAP estimation for over 900 models. Moreover, the environment variation is larger in telephone channel and MLLR can provide the adaptation for environmental changes.

3.5. Conclusions

All recognition results are improved after applying adaptation. It shows the model adaptation can significantly improve recognition performance. We also observed that the MLLR adaptation outperforms the MAP adaptation in our experiments. Therefore, we will use MLLR as our basic adaptation technique. In the MLLR adaptation, LMS MLLR is more robust to the noisy acoustic model and more computation-saving than standard MLLR does. Furthermore, we found that the optimal numbers of transformation in three tasks are different. It is 12 in Task 2 while it is 8 in Tasks 1 and 3.

By concluding the above observations, we will use the LMS MLLR and the optimal number of transformation in different tasks as our experiments setting.

References

[1] J. L. Gauvain, C. H. Lee, "Maximum a Posterior Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. SAP, Vol. 2, No. 2 , pp. 291 –298, April 1994.

[2] C. L. Leggetter, P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continues Density HMMs," Computer Speech and Language, No. 9, pp. 171-185, 1995.

[3] F. Wallhoff, D Willett, G. Rigoll, "Frame-Discriminative and Confidence-Driven Adaptation for LVSCR", ICASSP'00, pp. 1835-1838, 2000.

[4] D. Jiang, R. Zhao, "Speaker Normalization Based on the Generalized Time-Frequency Representation and Mellin Transform," ICSP2000, pp. 782-785, 2000.

[5] J. -I. Takahashi, S. Sagayama, "Fast Telephone Channel Adaptation Based on Vector Field Smoothing Technique," Interactive Voice Technology for Telecommunications Applications, 1994 IEEE Second Workshop, pp. 97-100, 1994.

[6] M. J. F. Gales, P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," Computer Speech and Language, No. 10, pp.249-264, 1996.

Chapter 4

Use of Confidence Measure for MLLR based Adaptation

4.1. Introduction to Confidence Measure

Confidence measure (CM) is used to indicate the reliability of a recognized speech segment. It has been proven to be useful in speech understanding, keyword spotting [1] and speech recognition [2]. For unsupervised adaptation of speech recognition systems, confidence measure acts as a control of the use of adaptation data [3][4]. Each utterance can be assigned a confidence score. Utterances that are recognized with higher confidence scores are more preferably utilized for adaptation. More precisely, if the confidence score of an utterance is above a certain threshold, it will be used in adaptation. An optimal threshold has to be determined to reach a good balance between the amount of usable data and confidence level.

Confidence measure can be computed at different levels of speech segment, depending on the intended applications. Word-level confidence measure is useful for keyword spotting while phrase-level measure is good for speech understanding. However, for model adaptation techniques like MLLR, Gaussian mixtures are the

target entities to be adjusted. In this case, confidence measure needs to indicate the mismatch at the Gaussian component level. In this regard, the reliability of a model contains more valuable information to the adaptation than that of a word. In this chapter, we will describe the use of model-level confidence measure. In addition, we are going to incorporate the information about the confusion between models into the estimation of confidence measure so as to reduce the effect of pronunciation variation.

4.2. Confidence Measure Based on Word Density

There are different ways of computing confidence measure from the output of a speech recognition system. They are based on acoustic scores, language model scores, length of words, frequency of word occurrences, etc. One of the most common approaches is based on weighted word density [1]. It uses not only the acoustic model and language model information but also the frequency of occurrences of a word in the N-best hypotheses. The higher the occurrence frequency of the word, the higher the reliability of this word hypothesis.

The ratio of the occurrence frequency of a hypothesized word to the total number of hypotheses is a good indicator of confidence. Each hypothesized word is weighted by the sentence score. The confidence score C_w of word w in hypothesis h is defined as,

$$C_w = \frac{\sum_{r \in E(w,h)} Q(S_r)}{\sum_{l=1}^{N} Q(S_l)} \qquad \text{eq. (4-1)}$$

$$Q(S_r) = P(S_r)^{\gamma} P(O \mid S_r)$$

where $Q(S_r)$ is the total path score, which is composed of the acoustic model probability $P(O|S_r)$ and the language model probability $P(S_r)$. The parameter γ is the acoustic model scaling factor.

$E(w,h)$ contains the indices of a set of hypotheses. Each of these hypotheses contains the word w at a time period that overlaps with the occurrence of w in h. The range of confidence score is between 0 and 1.

Such a confidence score is computed for each word occurrence in each input utterance. That is, the same word occurring in different sentence hypotheses would be treated separately and may have different confidence scores.

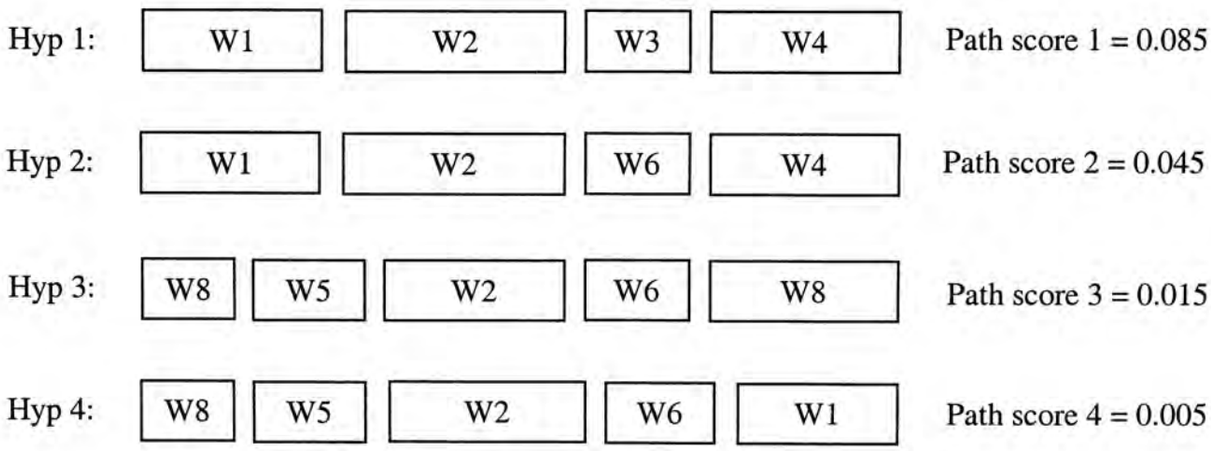


Figure 4-1: Word graph of the 4-best hypotheses

The above calculation can be explained with the example as shown in Figure 4-1. It shows the best 4 hypotheses, which contain totally 14 word occurrences. Each word occurrence is assigned a confidence score. For example, the confidence score of the word W1 in Hyp1 is computed as, according to eq. (4-1),

$$C_{w1(hyp1)} = \frac{\text{path score1} + \text{path score2}}{\text{path score1} + \text{path score2} + \text{path score3} + \text{path score4}}$$

$$= \frac{0.085 + 0.045}{0.085 + 0.045 + 0.015 + 0.005}$$

$$= 0.867$$

Since W1 is also found in Hyp 2, the numerator is the summation of path score 1 and path score 2. The denominator sums up all path scores. Although W1 is also included in Hyp 4, it does not appear in the same period as that in Hyp 1. Thus it is not included in the numerator.

Word density provides useful information about the reliability of a hypothesized word occurrence. This concept can be expanded to other speech segments of different sizes.

4.3. Model-level confidence measure

Doubtlessly, avoiding the use of erroneous recognition output is important for unsupervised adaptation. We need to know at which level the errors actually occur would have the most effect on adaptation, so that the useful information is better extracted and, at the same time, undesirable errors are reduced.

The basic unit to be adapted by the MLLR algorithm is Gaussian mixture. The recognition errors are caused by the mismatch at the Gaussian level. However, it is not straightforward to obtain the confidence measure for a Gaussian component. In previous research [3], word-level confidence measures were used as a substitution. However, using word-level confidence measures may cause the removal of not only the

errors but also useful data. For example, if there are three different words in three hypotheses, which all share the same model, this model should be regarded as being reliable and useful for adaptation. However, the word-level confidence score might not be high enough to pass the threshold because the word density is low.

We propose to use model-level confidence measure [5]. The benefit will be illustrated as in Figure 4-3 to Figure 4-5. The goal is to retrain more correctly recognized data for adaptation by providing a finer measurement of confidence.

Since the occurrence frequency is believed to provide useful information for unsupervised adaptation, it will be used as the basis for model-level confidence measure but the focused unit is model instead of word,

$$C_M = \frac{\sum_{r \in E(M,h)} Q(S_r)}{\sum_{l=1}^{N} Q(S_l)} \qquad \text{eq. (4-2)}$$

where $Q(S_r)$ is the path score. $E(M,h)$ contains the indices of a set of hypotheses. Each of these hypotheses contains the model M at a time period that overlaps with the occurrence of M in hypothesis h.

As described in Chapter 2, the speech recognition system for Cantonese uses Initials and Finals as the basic modeling unit. In the following discussion, model-level confidence measure will refer to that evaluated for each occurrence of Initial or Final. Both the context-independent base-IF and context-dependent biphone units will be investigated.

There are 956 biphone units and only 81 base-IFs being modeled. We call the respective confidence measures as biphone-based and baseIF-based confidence measures.

4.4. Integrating Confusion Information into Confidence Measure

Confusion matrix provides correlation information between speech units. It tells which units are confusing and how severe the confusion is. Confusion matrix has been found useful in pattern matching and pronunciation modeling [6].

In our research, confusion matrix is used to indicate the pronunciation variation of Cantonese syllables. Pronunciation varies with different speaking styles, accents and co-articulation. Sometimes, this variation may cause a phoneme to be mis-recognized as another phoneme. This mismatch would introduce recognition errors. Since the word-density based confidence measure is based on the occurrence frequency in the N-Best hypotheses, these recognition errors tend to degrade the estimation accuracy of confidence score. Therefore, by incorporating confusion matrix into the estimation of confidence measure, the effect from pronunciation variation can be alleviated. Here is the example to explain the incorporation of confusion matrix.

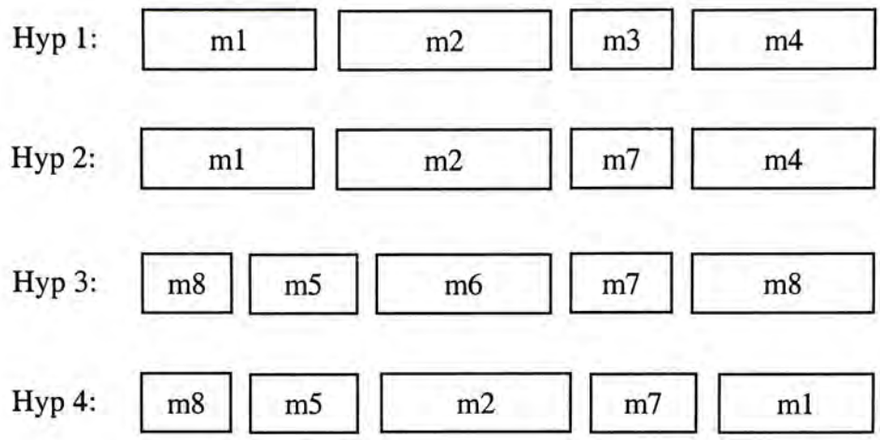


Figure 4-2: Example of the incorporation of confusion matrix

Figure 4-2 shows the model sequences of 4 recognition hypotheses. The model m2 in Hyp 1 and 4 are overlapped in time with m6 in Hyp 2. Suppose m2 is easily confused with m6. The presence of m6 is probably caused by pronunciation variation of m2 and in this case, m2 should deserve a higher confidence score.

It is required that the confusion matrix provides the confusing information between the acoustic models. Considering biphones as base-IF with the information of its right context, the confusion matrix is essentially similar to a context-dependent pronunciation variation dictionary. The following table shows parts of such a confusion matrix.

Baseform model M_B	Surface form model M_S	Occurrence Probability (%) $P(M_S\|M_B)$
I_g+F_ai	*I_g+F_ai*	91.66
I_g+F_ai	*I_g+F_ei*	4.17
I_g+F_ai	*I_z+F_ai*	4.17

Table 4-1: Part of the confusion matrix for biphone "*I_g+F_ai*"

The baseform model is the canonical model derived from the expected pronunciation while the surface form model is the actual model suggested by the recognizer. The occurrence probability indicated how often the baseform model is mapped to a particular surface form model.

We attempt to integrate such confusion information into the computation of model-level confidence measures. If both M_B and M_S are found in the N-best hypotheses during the same period of time, the confidence measure of the surface form model will be made contributive towards that of the baseform model. The modified confidence measure C_{M_B}' is re-computed as,

$$C_{M_B}' = C_{M_B} \times P(M_B \mid M_B) + \sum_{K \in G(M_{Sl},h)} [C_K \times P(M_{Sl} \mid M_B)] \qquad \text{eq. (4-3)}$$

where C_{M_B} and C_{M_S} are the original confidence measure of M_B and M_S respectively. M_{Sl} is the lth surface form model. G(M_{Sl},h) is the set of models which contains the surface from model M_{Sl} with overlapping time period with model M_B in hypothesis h.

In eq. (4-3), C_{M_B} and C_{M_S} are scaled by the occurrence probabilities of P_{Sl}. Thus the effect of C_{M_S} depends on the degree of confusion.

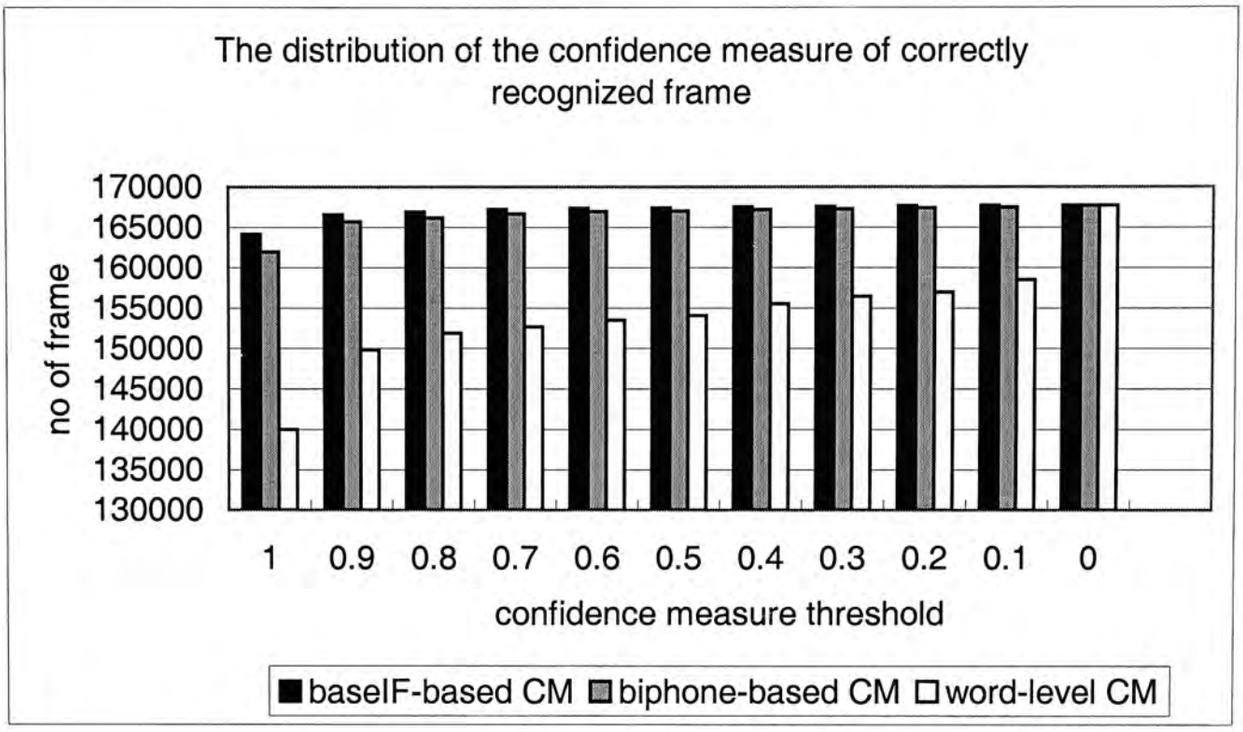
4.5. Adaptation Data Distributions in Different Confidence Measures

Confidence measure is used for data selection. If the confidence measure of a word or a model is above a certain threshold, it will be used in the adaptation. The amount of data

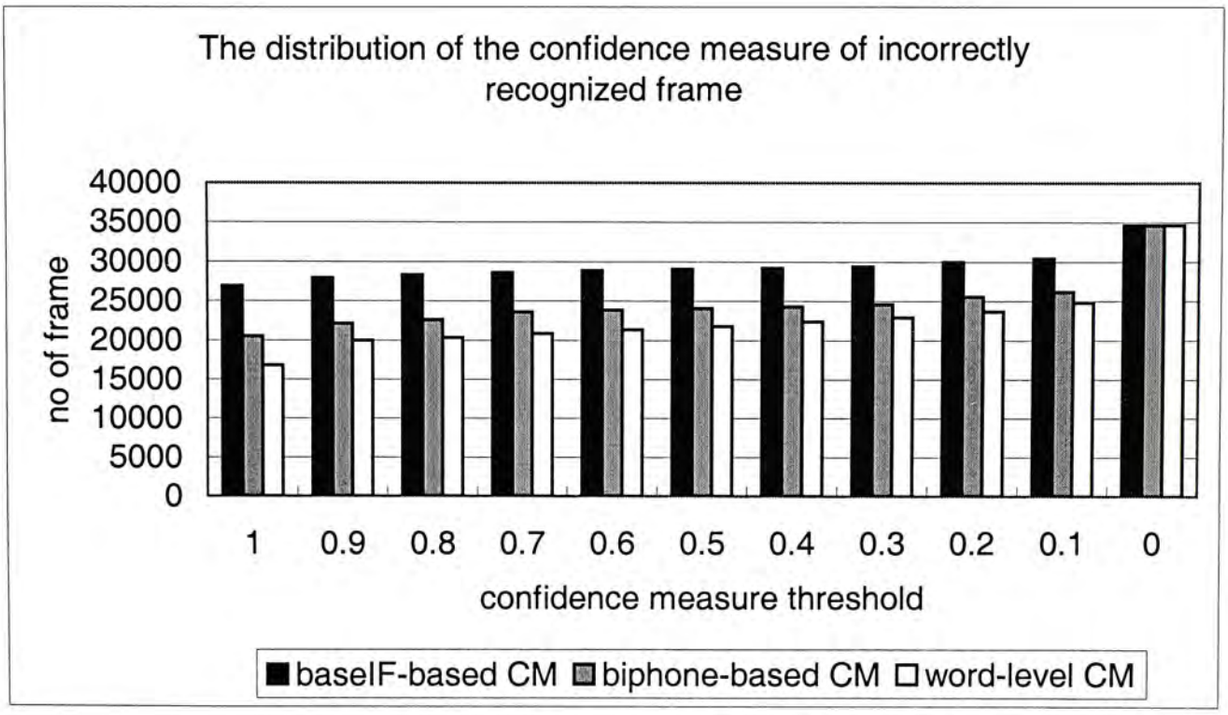
increases with the threshold decreasing. All data would be used in adaptation when the threshold is set to zero. If the threshold is too low, it would lose its function of filtering errors. Therefore, we need to evaluate the amount of correctly and incorrectly recognized frames against different thresholds when different confidence measure methods are applied.

We will analyze the distribution of adaptation data in all three tasks that we described in Chapter 3. All adaptation and testing utterances are used. The maximum number of hypotheses is set to be 50. All data would be included in the calculation of confidence measures. And the confidence measure of data in the first 10-best hypotheses will be evaluated. We would compare the recognized model in each hypothesis with the actual model frame by frame.

Task 1



(a)



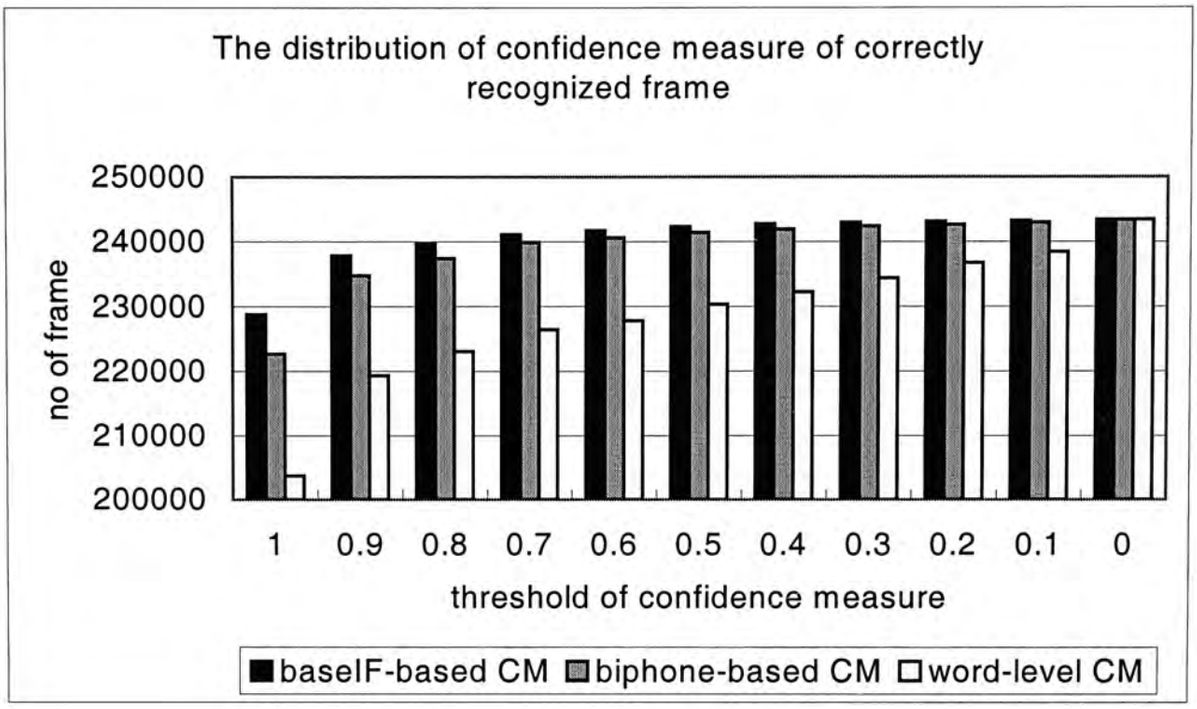
(b)

Figure 4-3: (a) The amount of correctly recognized data for which the confidence scores are above or equal to difference thresholds in Task 1. (b) The amount of incorrectly recognized data for which the confidence scores are above or equal to the thresholds in Task 1

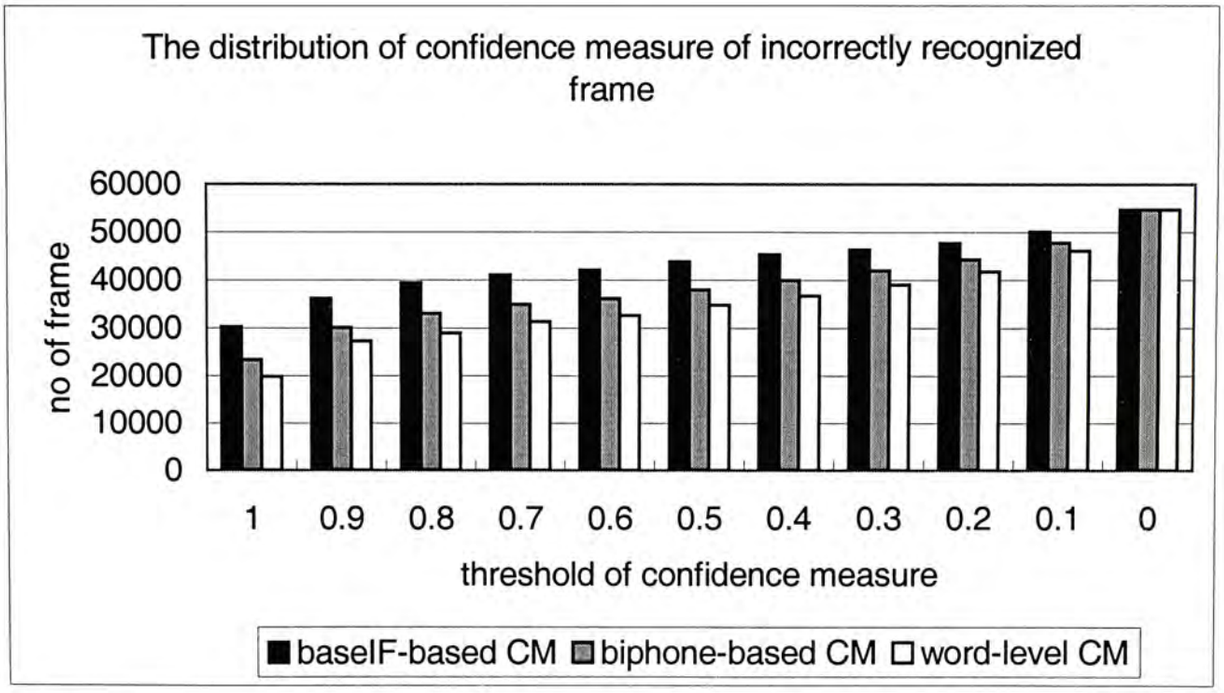
In these figures, we can see the amount of incorrectly and correctly recognized data. The ratio of incorrectly and correctly recognized data is about 1:5, which is directly related to the baseline WER. Since the recognition system has a low WER in Task 1, the ratio is relatively small. Comparing with the word-level confidence measure, the model-level method assigns more correctly recognized data with high confidence scores. When threshold is 0.5, nearly 100% of correctly recognized data are retained for adaptation. Compared with the word-level method, there are 7% more good adaptation data to be used. However, more errors are included at the same time.

It is observed that most of the data are assigned the two extreme confidence scores, i.e. 0 or 1. Most of the correctly recognized data are scored as 1. Even if the threshold is lowered, the amount of good adaptation data would not increase much. Moreover, the amount of incorrect data increases greatly in the Figure 4-3(b) when the threshold is moving from 0.1 to 0.

Task 2



(a)



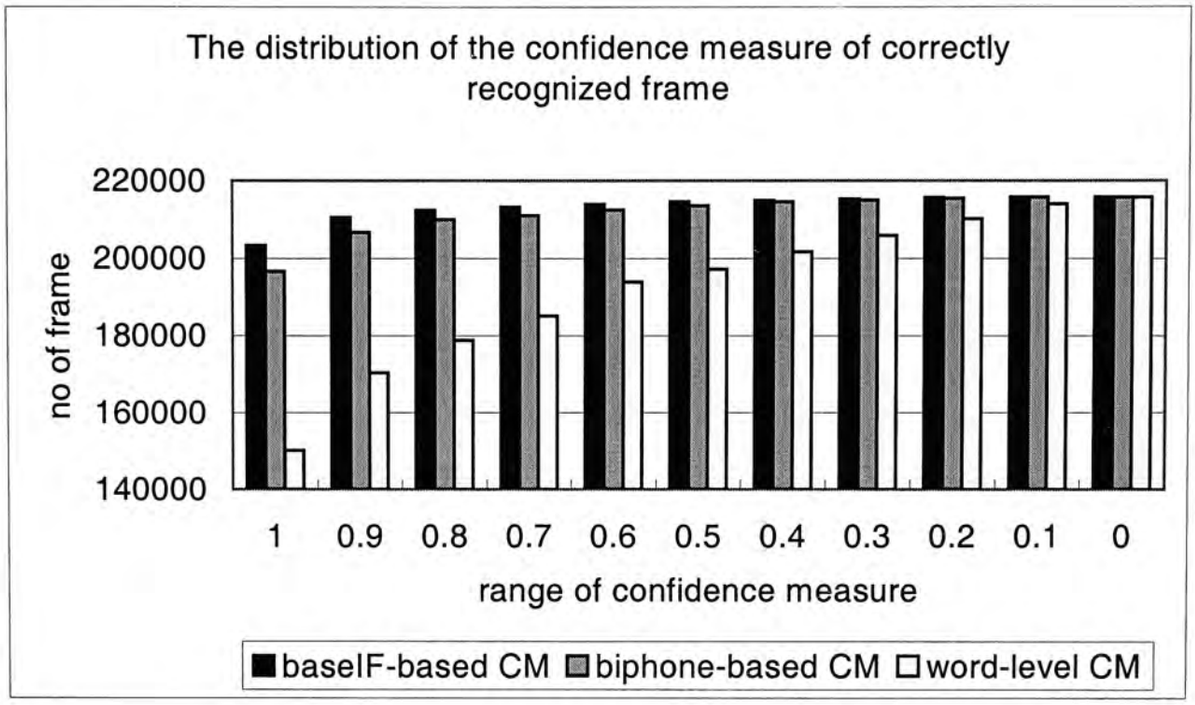
(b)

Figure 4-4: (a) The amount of correctly recognized data above or equal to the threshold of confidence measure in Task 2. (b) The amount of incorrectly recognized

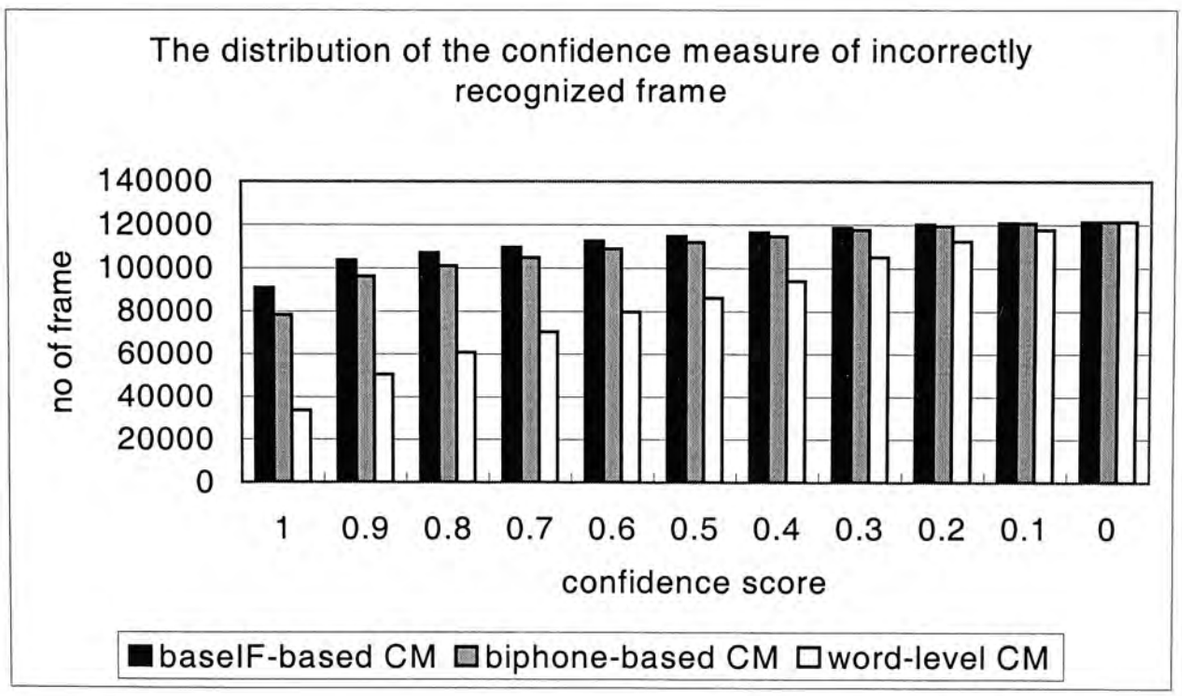
data above or equal to the threshold of confidence measure in Task 2

Similar to that in Task 1, the model-level confidence measure can retain more correctly recognized data. However, the distribution of data is more even in this task. The confidence score of correctly recognized data does not concentrate around 1 and that of incorrectly recognized data does not concentrate at 0. This is probably due to a higher WER in Task 2. The estimation of confidence measure is less accurate.

Task 3



(a)



(b)

Figure 4-5: (a) The amount of correctly recognized data above or equal to the threshold of confidence measure in Task 3. (b) The amount of incorrectly recognized data above or equal to the threshold of confidence measure in Task 3

The ratio of incorrectly to correctly recognized data is 1:2 in Task 3. The baseIF-based confidence measure can weight most data to a higher confidence score while the word-level confidence measure can weight less data. Either the distribution of correctly and incorrectly recognized data are even.

It is observed that model-level confidence measures tend to retain more data than those computed for higher-level speech segments. Thus, they are more suitable for unsupervised adaptation. Nevertheless, model-level confidence measures cause more false acceptance than the word-level ones. This may partially offset the benefit from the increase of useful data. We expect that the contribution of useful data will overwhelm the effect of the erroneous data. This will be verified by the experiment results in next chapter.

References

[1] M. Weintraub, "LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting," ICASSP1995, pp. 297-300, 1995.

[2] F. Wessel, R. Schluter, K. Macherey, H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Volume: 9 Issue: 3, 2001.

[3] F. Wallhoff, D Willett, G. Rigoll, "Frame-Discriminative and Confidence-Driven Adaptation for LVSCR", ICASSP'00, pp. 1835-1838, 2000.

[4] D. Charlet, "Confidence-Measure-Driven Unsupervised Incremental Adaptation for HMM-Based Speech Recognition", ASSP 2001, vol 1, pp. 357-360, 2001.

[5] K. Y. Kwan, T. Lee, C. Yang, "Unsupervised N-Best Based Model Adaptation Using Model-Level Confidence Measure", ICSLP 2002', (paper accepted for publication)

[6] M. Liu, B. Xu, T. Hunng, Y. Deng, C. Li, "Mandarin accent adaptation based on context-independent /content-dependent pronunciation modeling", ICASSP'00, pp. 1025-1028, 2000.

Chapter 5

Experimental Results and Analysis

There are many factors affecting the effectiveness of model adaptation, e.g. the amount of adaptation data, reliability of the data and the WER of the recognizer. Reliability of adaptation data in unsupervised adaptation depends on the recognition result. In our work, we try to select the data with high reliability by using confidence measure. The estimation of confidence measure was introduced in the previous chapter, based on the N-best recognition framework.

The effectiveness of using confidence measures for adaptation will be evaluated by simulation experiment in this chapter. Firstly, we will get a general idea of how the standard MLLR algorithm performs in different tasks. The improvements attained with supervised and unsupervised adaptation are compared. Secondly, we will evaluate the effectiveness of cheated confidence measure, which may also be called perfect confidence measure. In this case, no recognition error is allowed for the adaptation. The hypotheses are compared with known model-level transcription. The selection of data is done manually by identifying the hypothesized models that are matched with transcription. The cheated confidence measure should indicate the maximum contribution of confidence measure to unsupervised adaptation. Finally, the two

proposed confidence measure, namely the model-level confidence measure and the confidence measure incorporating confusion information are investigated.

Recognition experiments are performed for the three tasks described in Chapter 3.

Task 1: Domain-specific microphone speech recognition

Task 2: Domain-specific telephone speech recognition

Task 3: General-domain telephone speech recognition

Table 5-1 gives a summary of the experimental results.

5.1. Supervised Adaptation

		Baseline	Supervised adaptation	No confidence measure	Cheated confidence measure
Task 1	**WER(%)**	13.05	8.66	9.14	9.93
	Relative improvement (%)	--	33.64	29.96	23.91
Task 2	**WER (%)**	18.89	14.47	16.03	14.41
	Relative improvement (%)	--	23.4	15.14	23.72
Task 3	**WER(%)**	40.88	35.35	35.72	35.2
	Relative improvement (%)	--	13.53	12.62	13.90

Table 5-1: The WER(%) and relative improvement(%) in three tasks.

In supervised adaptation, the true transcription is used for adaptation. Thus there is no recognition error that affects adaptation. The performance of supervised adaptation is considered to indicate the upper bound of all techniques being investigated. By comparing the results among different tasks, we try to understand how the effectiveness of model adaptation depends on the acoustic environments and the task domains.

As shown in Table 5-1, supervised model adaptation gives the most noticeable improvement in Task 1. The relative improvement is 33.64%. Although the application domains of Task 1 and 2 are the same, the relative improvement for Task 2 is considerably smaller than for Task 1. This is apparently due to their different channel and acoustic conditions. Task 1 deals with microphone speech and Task 2 is for telephone speech. Telephone speech is acquired under a much more adverse environment, in which the noise may cause large acoustic variation. The model adaptation technique can not do much in terms of handling acoustic variation.

On the other hand, the relative improvement for Task 2 is much more significant than that for Task 3. The recognition systems in both tasks deal with telephone speech but they work in different application domains. Task 3 involves a much large vocabulary size and a wide variety of contents. This results in relatively loose linguistic constraints for the recognition process. Reduced restriction from the language model and the lexicon tends to lower the recognition accuracy. In this case, acoustic mismatch is not the major cause of recognition errors so that the effectiveness of model adaptation is not very significant in Task 3.

5.2. Cheated Confidence Measure

The performance improvement attained with cheated confidence measure indicates the maximum contribution that confidence measure can provide.

Table 5-1 shows that, for Task 2, using cheated confidence measure is just as good as supervised adaptation. The relative improvement is 23% while that without using any confidence measure is just 15%. Filtering incorrectly recognized data is undoubtedly helpful to adaptation and the use of confidence measure contributes positively to adaptation.

We can also observe slightly improvement in Task 3. Although the WER is just reduced a little, it approaches to that of supervised adaptation. The limitation of the performance in this case is similar to that we explained in the last section "Supervised Adaptation".

It is observed that the use of cheated confidence measure in Task 1 makes the system perform even worse than that before the integration. This result is deviated from what we expected. Since the cheated confidence measure has filtered all the model-level incorrectly recognized data, we hypothesize that there may be some useful information in these incorrectly recognized data and their absence causes performance degradation. A detailed analysis is given in the next paragraph.

Since we label and remove incorrectly recognized data at model-level, these data may be correct at state or class level. When the hypothesized state is matched with the actual state transcription, the same Gaussian mixture is adapted though they belong to different HMM model. Moreover, since MLLR is a class-based adaptation,

distributions are first clustered into different classes and those in the same class share a transformation. The estimation of transformation is sensitive to the amount of adaptation data. More reliable transformation can be estimated with a larger amount of adaptation data. Therefore, if the model-level incorrectly recognized data are in the same class with the actual transcription of adaptation utterance, they may be useful in the estimation. These two factors can both give contribution to the adaptation.

As shown in Table 5-2, 38% of the model-level incorrectly recognized data have same regression classes with the actual transcription and 13% even match the exact states. These percentages are found to be much higher than the other two tasks. Nearly 40% of these removed data are considered to be useful for adaptation.

	Task 1	Task 2	Task 3
Correct State/ total incorrectly recognized data (%)	13.76	10.76	12.08
Correct Class/ total incorrectly recognized data (%)	38.25	33.12	32.89

Table 5-2: The ratio of correct state and class in the model-level incorrectly recognized data in three tasks

Moreover, the microphone speech is cleaner than the telephone speech, which contains fewer noise data. Therefore, even for the incorrectly recognized data, the negative effect is smaller.

5.3. Confidence Measures of Different Levels

Task 1

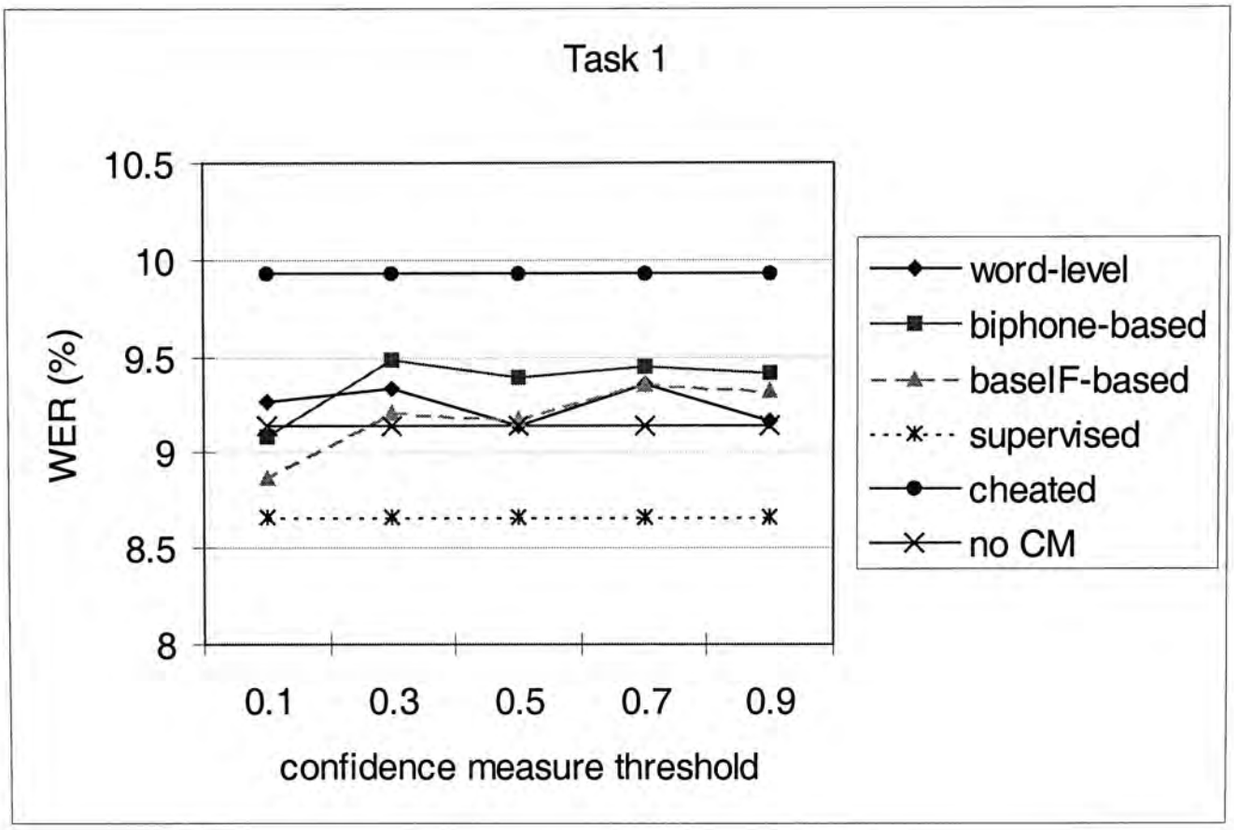


Figure 5-1: The WER (%) in Task 1 when confidence measures of different levels are integrated. "word-level", "biphone-based" and "baseIF-based" denote different levels of confidence measure. "no CM" denotes adaptation without using confidence measure. "supervised" demotes supervised adaptation. "cheated" denotes adaptation with cheated confidence measure.

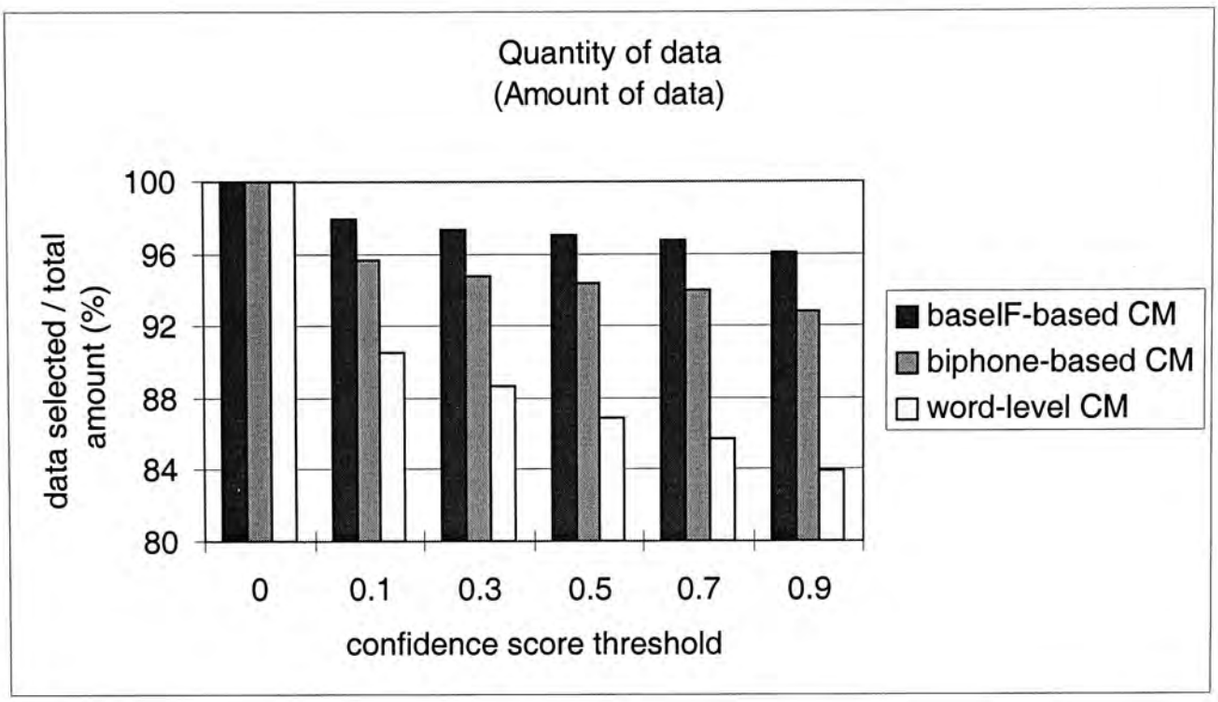


Figure 5-2: The percentage of data above or equal to the threshold of confidence measure so they are selected for adaptation

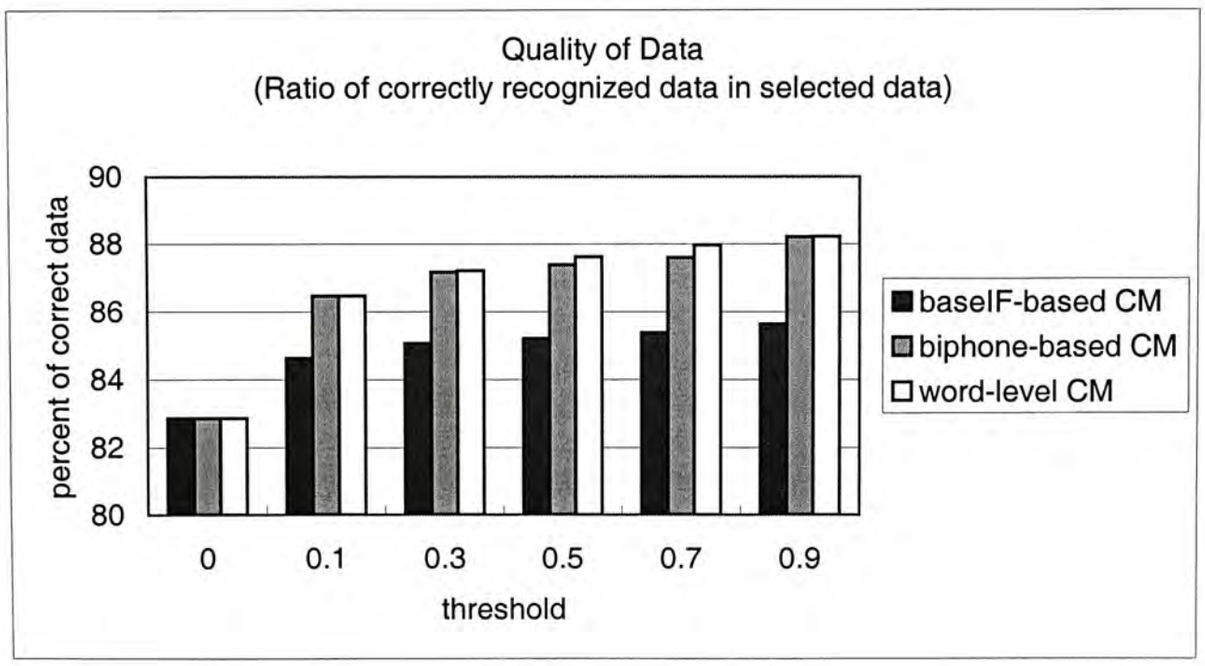


Figure 5-3: The ratio of correctly recognized data in the selected data (Quality of the selected data)

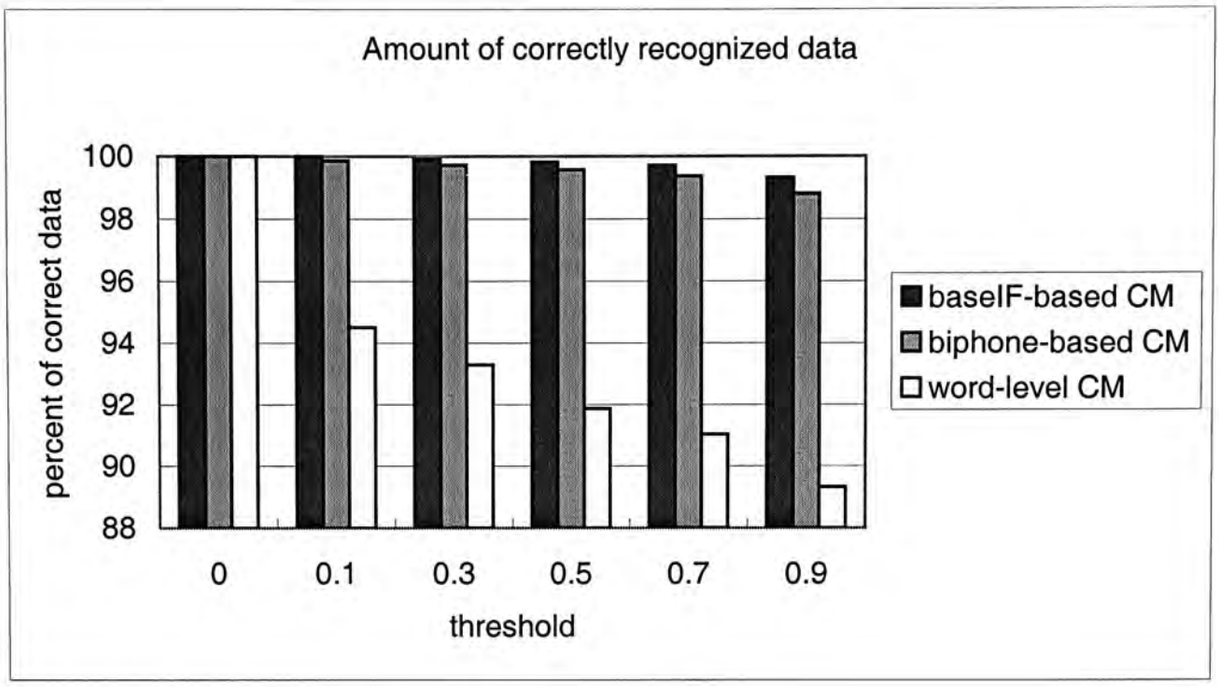


Figure 5-4: The amount of correctly recognized data

From Figure 5-1, it is shown that confidence measures do not benefit to adaptation almost in all cases. An exception occurs in model-level confidence measure when the threshold is equal to 0.1. The baseIF-based gives the best performance. And it is surprising that the WER is improved with the decrease in threshold in baseIF-based confidence measure. It is observed that over 99% of correctly recognized data for base-IF based case are selected when the threshold is 0.8 in Figure 5-4. Therefore, the decrease in threshold makes more incorrectly recognized data are selected whereas the amount of correctly recognized data is not affected much. As we mentioned in last section, the absence of incorrectly recognized data degrades the results and the reason is that some data are recognized incorrectly in model-level but correct in state or class level. Such characteristic makes them have contribution to adaptation still.

We can find the explanation for the exception of low WER for threshold 0.1. From Figure 5-5, we found that most of the incorrectly recognized data in the same class of actual transcription are selected and more than 10% of the data in the different classes

are filtered when threshold is 0.1. There are some incorrectly recognized data that cause error in adaptation but the amount is not much and they are usually weighted to a low confidence score. Furthermore, it is observed from Figure 5-3 that the quality of data decrease from 86% to 82% when the threshold is moving from 0.1 to 0. This significant drop makes the performance degradation when moving the threshold from 0.1 to 0.

Three confidence measures can successfully weight the correctly recognized data to high confidence score and the incorrect ones to low confidence score. The distribution of data is in two extreme confidence score. When the threshold is varied between 0.1 – 0.9, the quality and quantity of data are kept almost the same. Therefore, we can observe that the WER can still keep in a narrow range between 9-9.5% when we vary the threshold.

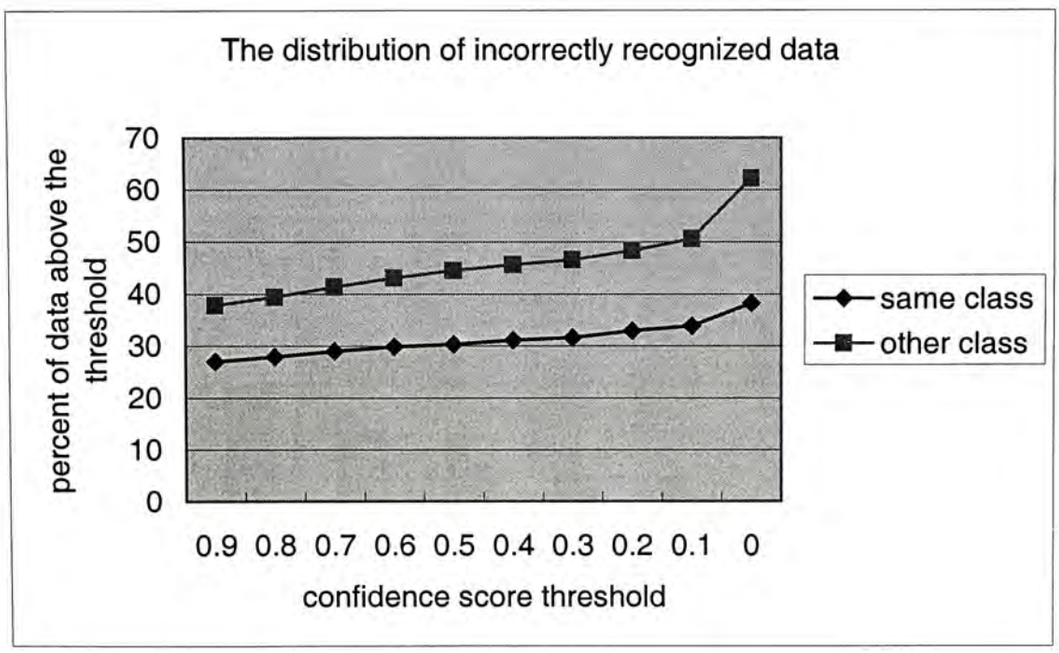


Figure 5-5: The distribution of incorrectly recognized data which are in the same regression class with the actual transcription alignment and those which are in the other classes.

Task 2

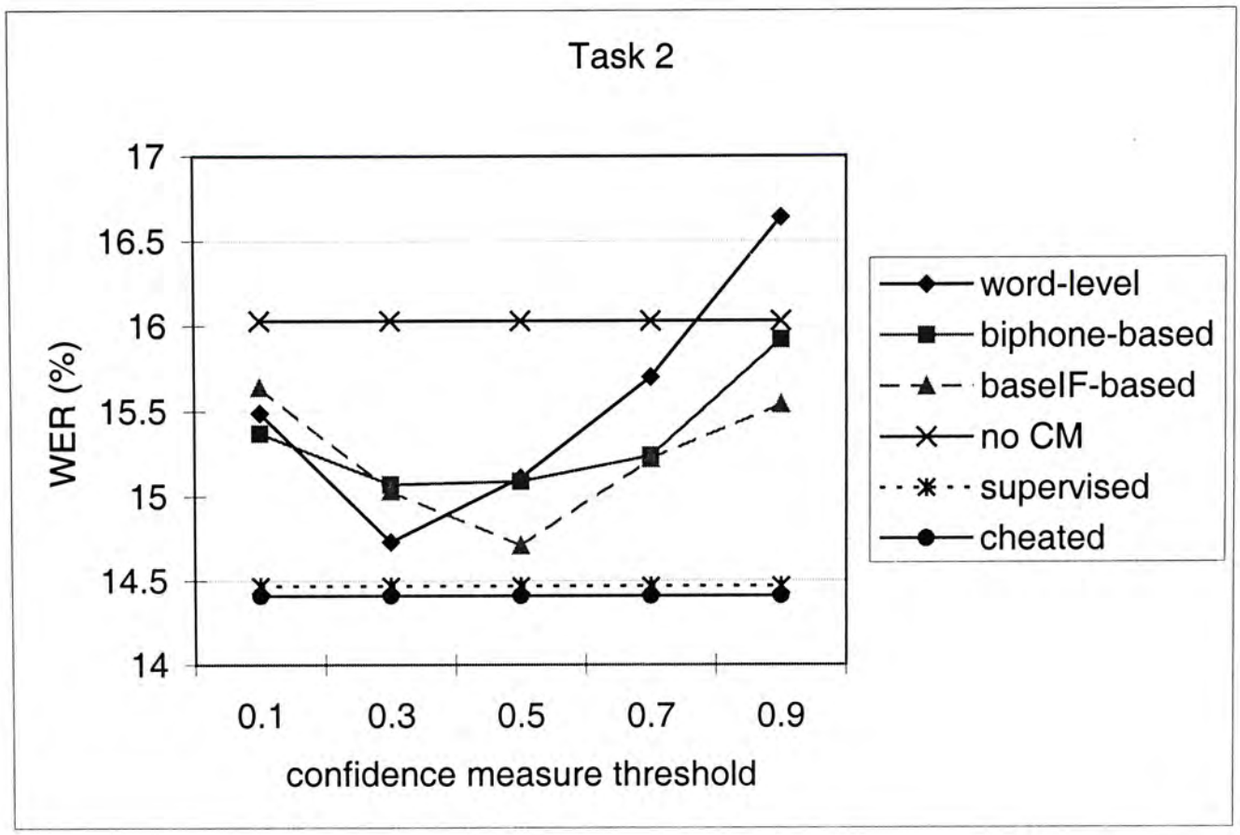


Figure 5-6: The WER (%) in Task 2 when confidence measures of different levels are integrated.

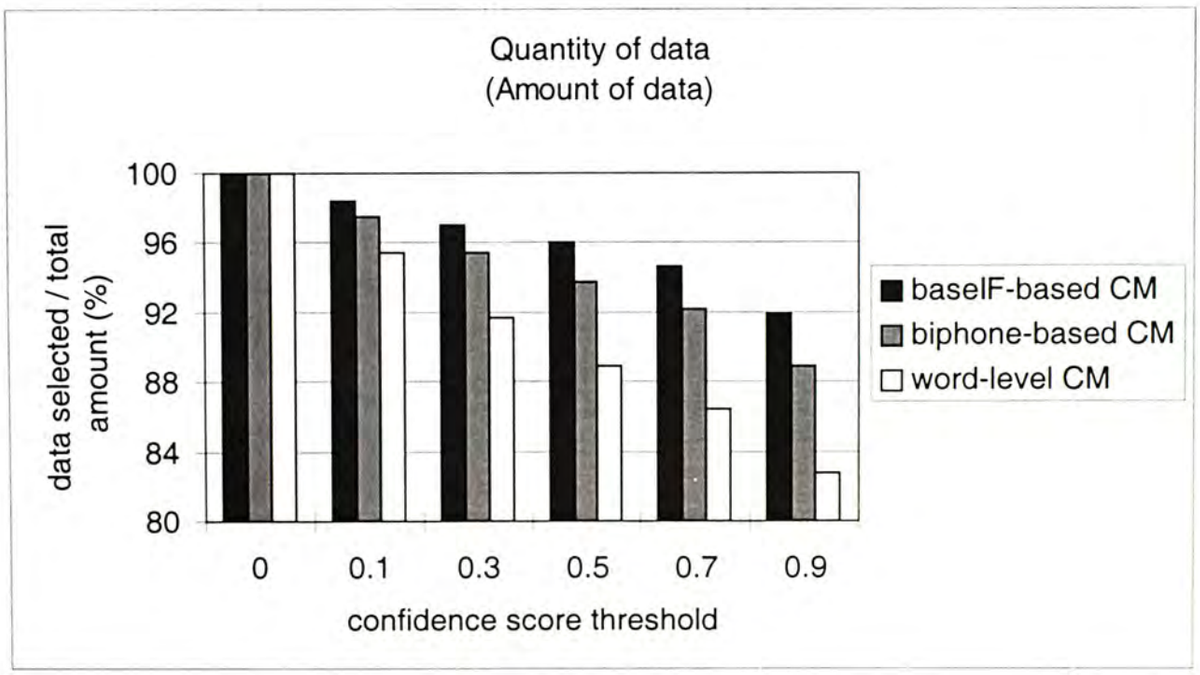


Figure 5-7: The Quantity of data which is represented by the amount of data

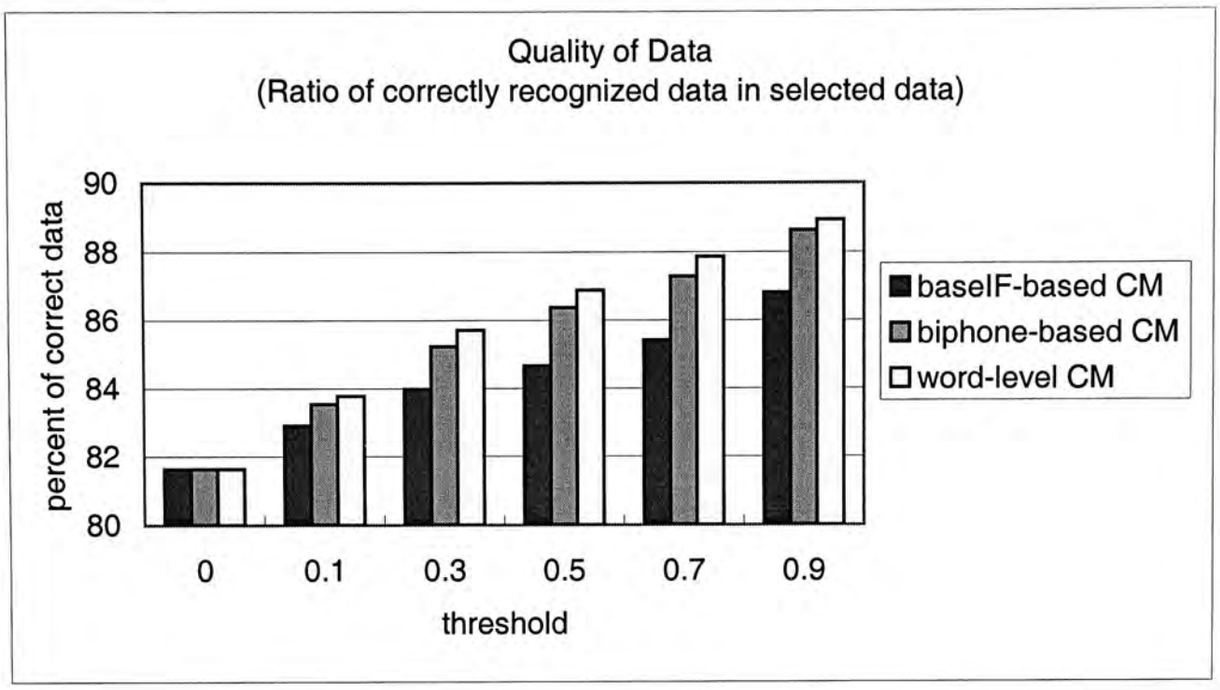


Figure 5-8: The quality of data which is determined by the ratio of correctly recognized data in selected data

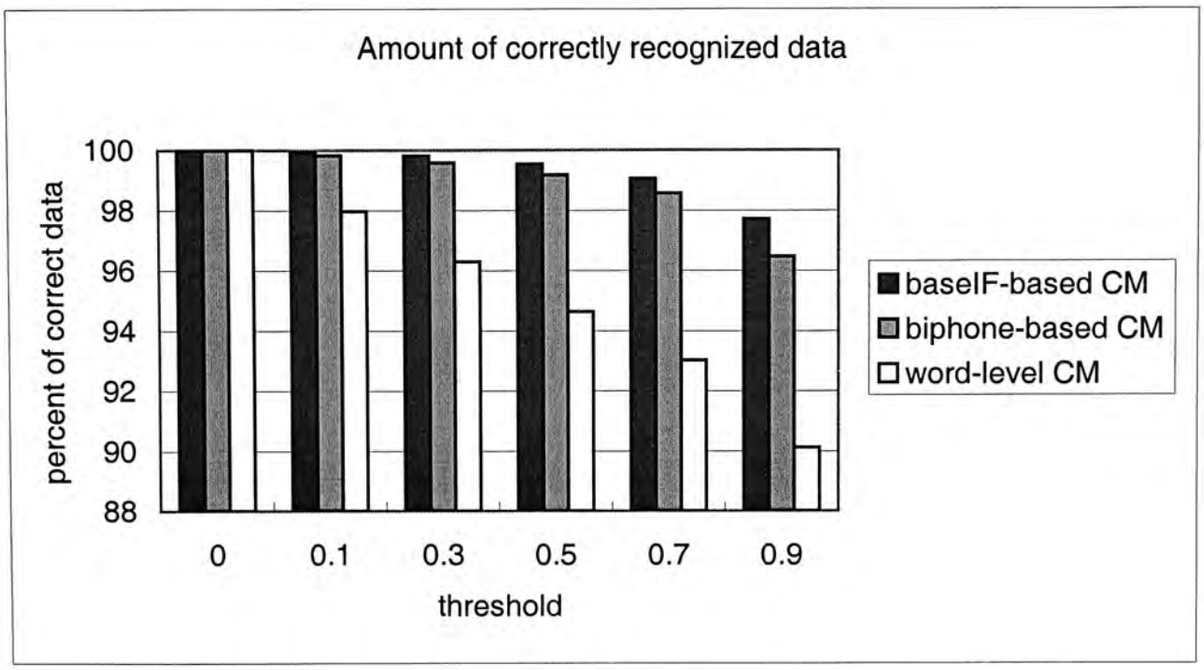


Figure 5-9: The amount of selected correctly recognized data out of all correctly recognized data

The results in this task are what we expect. The integration of confidence measure improves the recognition result. The optimal thresholds for different methods are found between 0.3 and 0.5. BaseIF-based confidence measure can even lower the WER to 14.71%, which is quite close to the lower bound given by supervised adaptation. Word-level confidence measure also gives a good result when the threshold is 0.5 but it is slightly worse than baseIF-based one. Moreover, in the word-level confidence measure, the change of WER with the varying threshold is larger than the other two methods. And the performance is even worse than the baseline when the threshold is 0.9. Since the word-level confidence measure cannot retain the correctly recognized data to high confidence score, the quantity of data are degraded a lot with high threshold. Based on the same reason, the optimal threshold of word-level case is lower than that in model-level cases.

The distributions of incorrectly and correctly recognized data in Task 2 are less extreme than that in Task 1. We can observe that the quantity of data increases with the decrease in threshold whereas the quality of data decreases.

Task 3

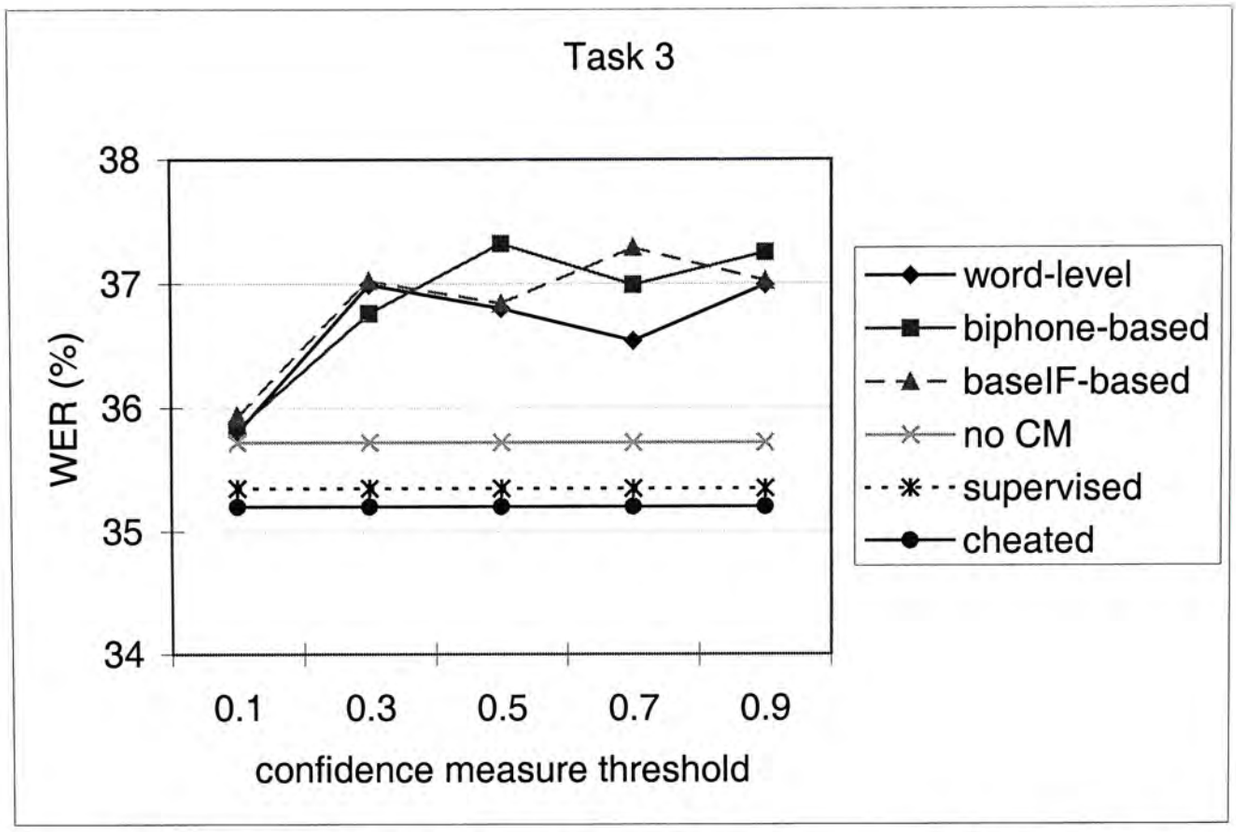


Figure 5-10: The WER (%) in Task 3 when confidence measures of different levels are integrated

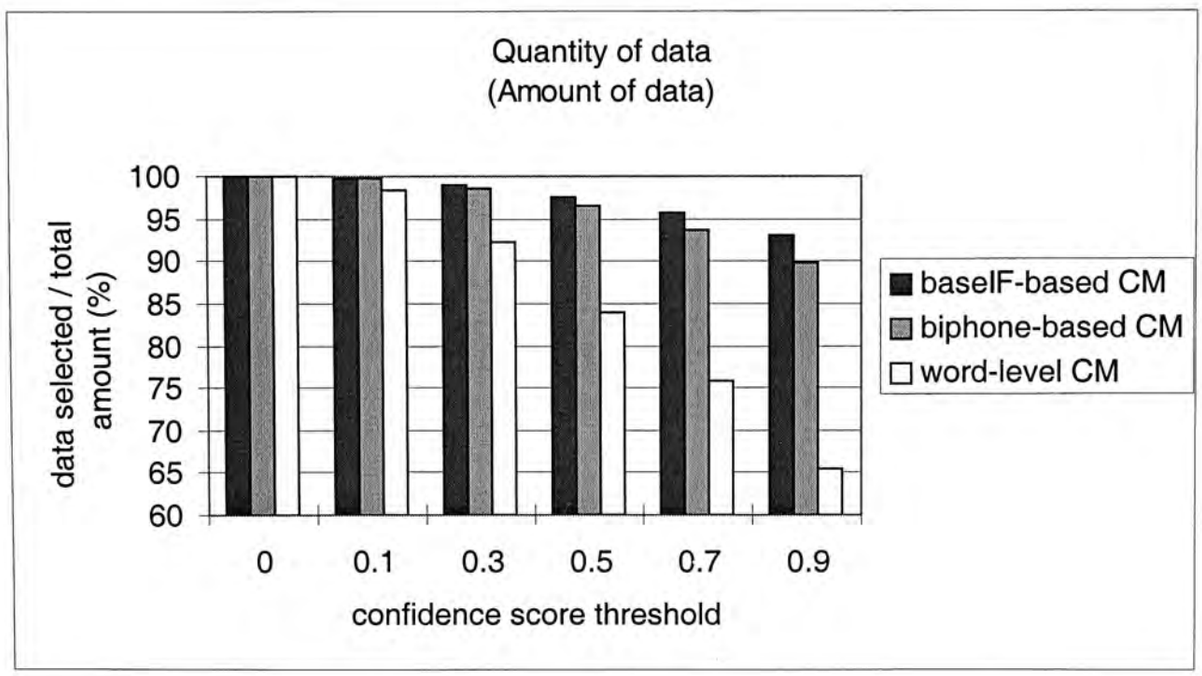


Figure 5-11: The quantity of data is determined by the percentage of data above the threshold

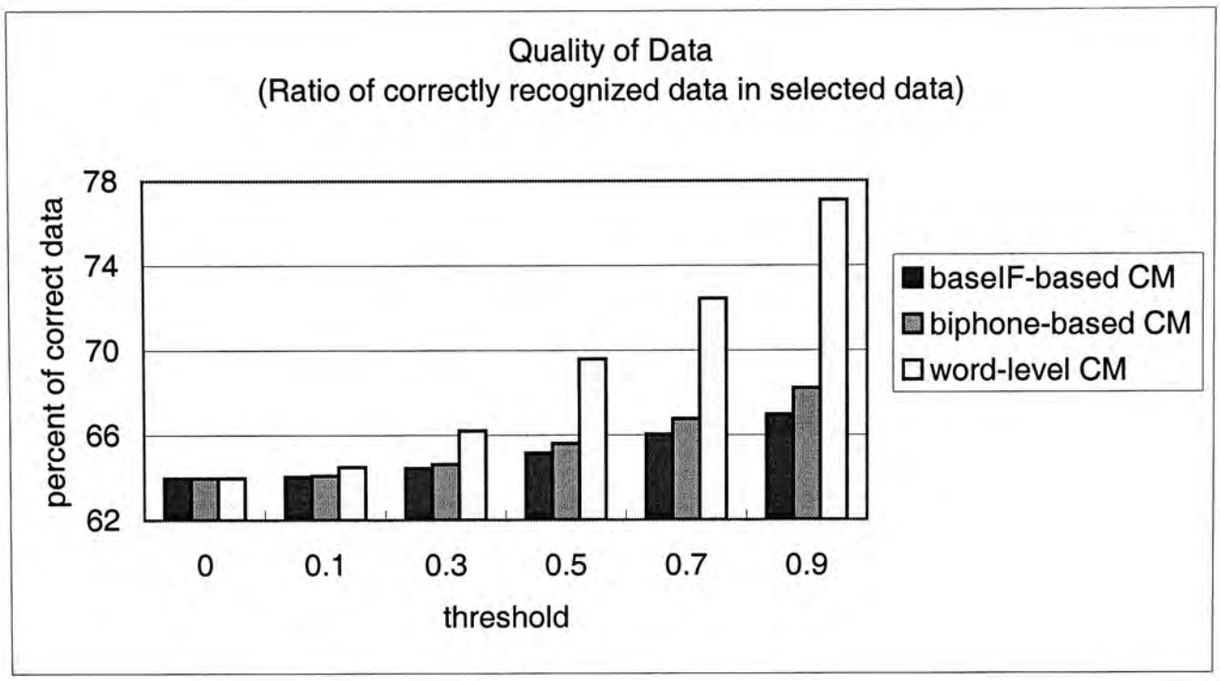


Figure 5-12: The quality of data which is determined by the ratio of correctly recognized data in selected data with different threshold

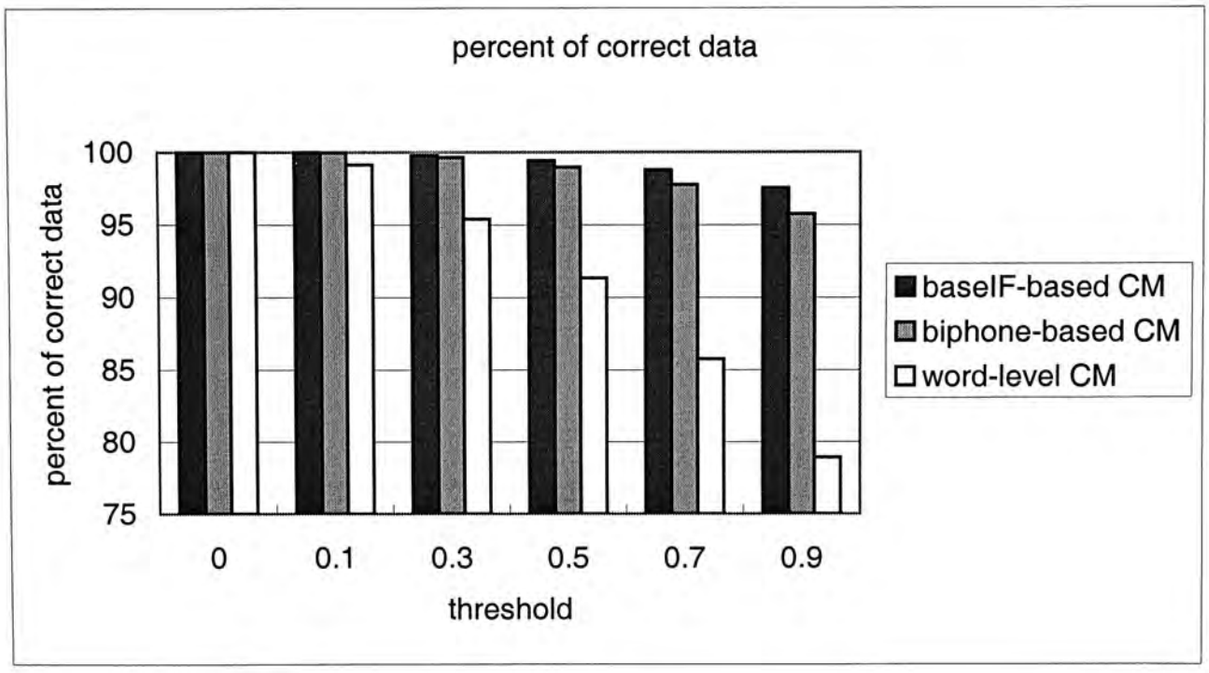


Figure 5-13: The percentage of correctly recognized data above or equal to the threshold

We were expecting the confidence measure would be useful in Task 3, given the improvement with cheated confidence measure. However, all of the confidence measures make the adaptation less effective. The trend of their performance with varying threshold is similar.

The high baseline WER is the main cause of unsatisfactory results. The algorithm of confidence measure in our research is based on the occurrence frequency. The estimation accuracy is significantly affected by the high WER. Therefore, it is observed that the model-level confidence measures do not select data with good quality in Figure 5-12. Although the word-level confidence measure can give a good data quality when high threshold is used, the amount of data become insufficient and affects the performance of adaptation. It is hard to have a good balance when the WER is high.

5.4. Incorporation of Confusion Matrix

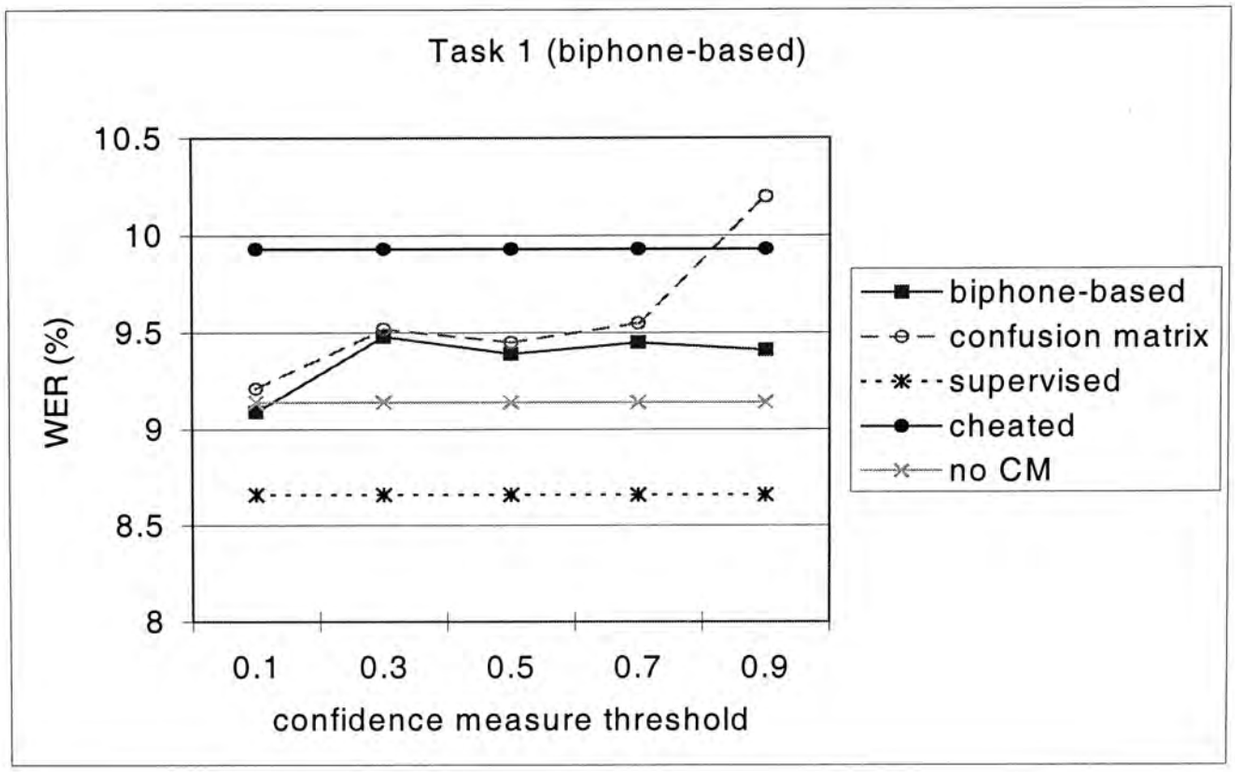


Figure 5-14: The WER(%) of biphone-based confidence measure incorporating with confusion matrix in Task 1

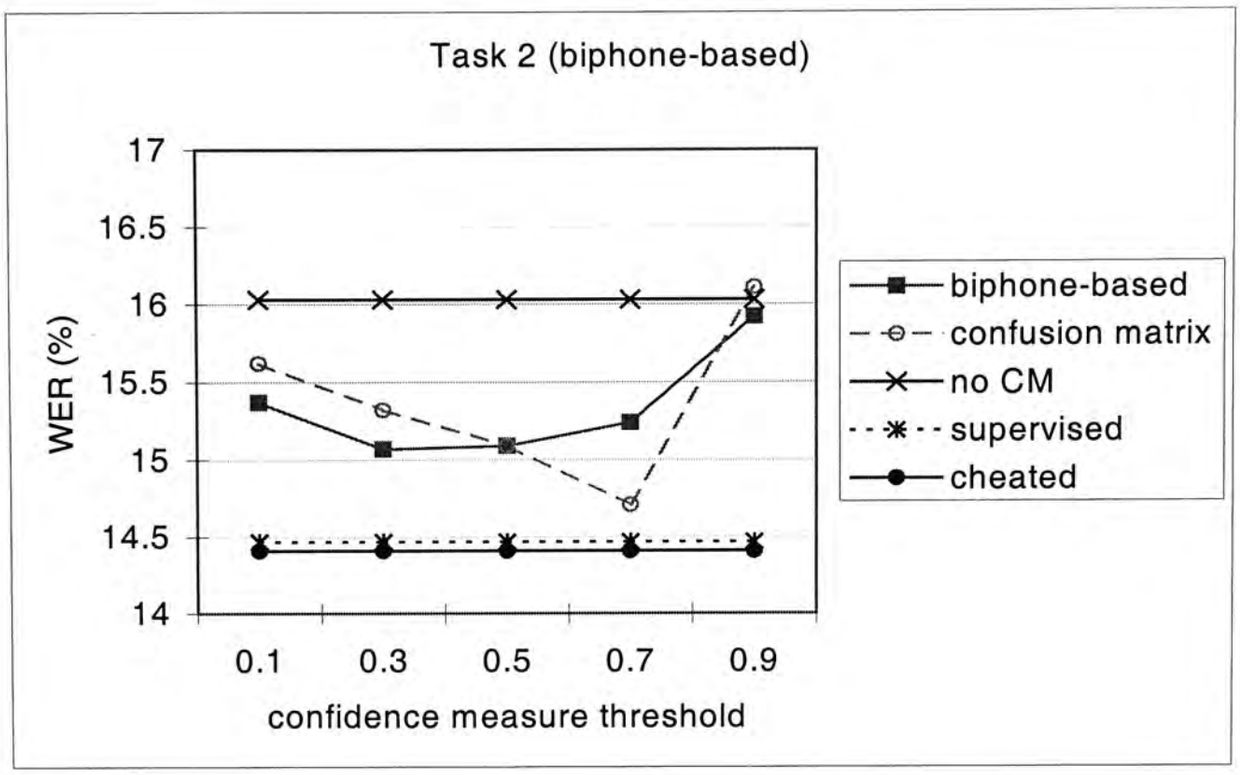


Figure 5-15: The WER(%) of biphone-based confidence measure incorporating with confusion matrix in Task 2

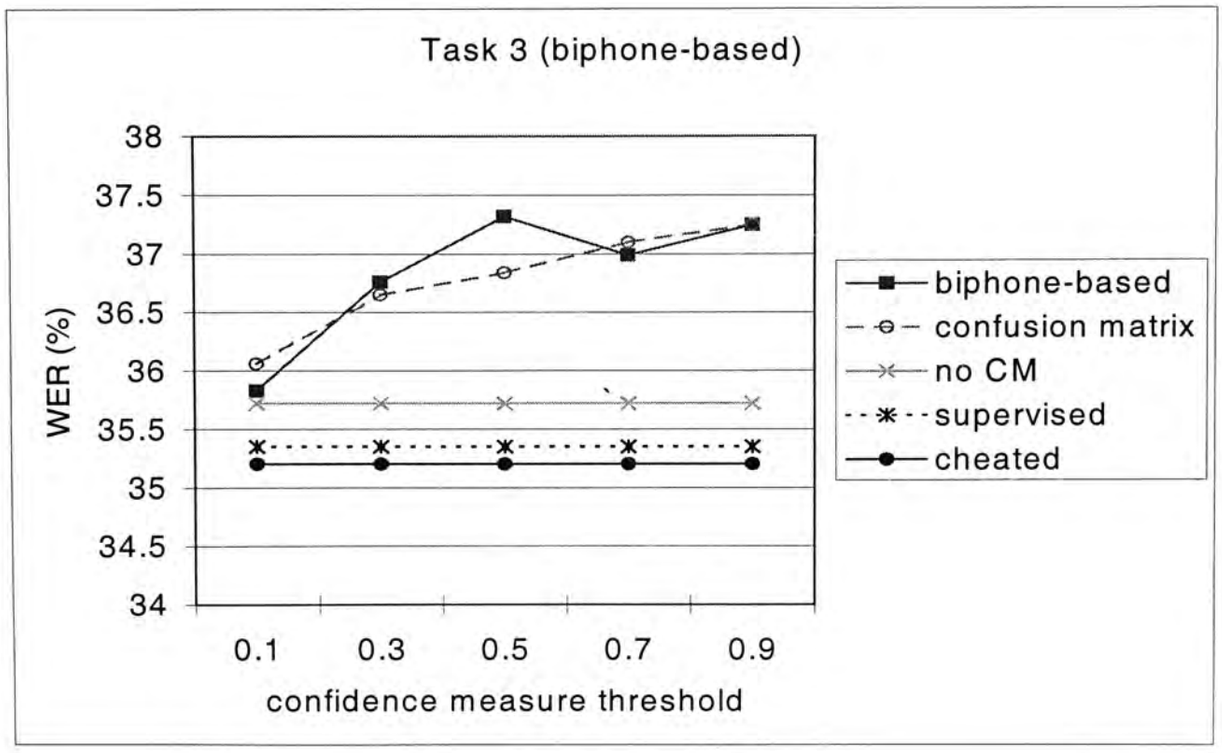


Figure 5-16: The WER(%) of biphone-based confidence measure incorporating with confusion matrix in Task 3

We can observe an improvement when we add the information of confusion matrix in the biphone-based confidence measure in Task 2. The WER can be as low as 14.71%. The improvement is 8% from no confidence measure applied. Therefore, the information of confusion matrix is useful to the estimation of confidence measure. Moreover, the optimal threshold is 0.7, which is higher than before incorporation. This means that the use of confusion information can move the correctly recognized data towards high confidence.

However, similar to the last section, confidence measure cannot give any improvement even the information of confusion matrix is added in Task 1 and 3. The accuracy is similar to the original confidence measure.

5.5. Conclusions

The use of confidence measure shows significant contribution in MLLR based adaptation for domain-specific telephone recognition system. One of the proposed methods, baseIF-based confidence measure, gives the greatest improvement. The WER can be reduced from 16.03% to 14.71%. Moreover, the incorporation of confusion matrix to biphone-based confidence measure can also improve the WER to 14.71%. Both methods can give significant results.

However, the channel and domain conditions significantly affect the performance of confidence measure. For Task 2, which is a microphone speech recognition system,

the removal of all model-level incorrectly recognized data degrades the adaptation performance. Some of these incorrectly recognized data contain useful information. However, it is also observed that some of them with low reliability are not desired in the adaptation. Therefore, filtering the unreliable data is necessary in this task too but more information should be added in the estimation of confidence measure, such as state and class information. In Task 3, the low baseline WER degrades the accuracy of reliability estimation. In order to tackle this problem, information that is more robust to recognition performance should be incorporated in the estimation.

Chapter 6

Conclusions

This research focuses primarily on unsupervised model adaptation for HMM based continuous speech recognition. We have investigated on the use of confidence measure to improve the effectiveness of model adaptation. For each adaptation utterance, the confidence measure is computed from the N-best output hypotheses and their associated path scores. The results are used to determine whether the relevant speech data should be utilized to guide the adjustment of HMM parameters or not.

MLLR has been adopted as the basic adaptation technique. Given the properties of the MLLR, model-level confidence measure is considered to be more appropriate than word-level and utterance-level ones. Therefore, a model-level confidence measure has been proposed. It facilitates the selection of adaptation data at frame level. Compared with the case of using word-level confidence measure, the amount of good adaptation data is increased by at least 5%.

The proposed use of model-level confidence measure is examined on adapting the acoustic models in three different tasks of continuous Cantonese speech recognition. The observations are summarized as follows.

1. Domain-specific telephone speech recognition system

The WER is reduced from 16.03% to 14.71% by using the proposed confidence measure to the adaptation. It is closed to the performance of supervised adaptation. One of the model-level confidence measure, baseIF-based, can give the best performance. When another model-level confidence measure, biphone-based, is incorporated with the confusion matrix, the same improvement is observed.

In this task, confidence measure is helpful to remove unreliable recognition output and to improve the adaptation. The incorporation of confusion matrix brings further improvement.

2. General domain telephone speech recognition system

When we apply the cheated confidence measure, a significant improvement can be observed. The WER is as low as that of supervised adaptation. Selecting reliable adaptation data seems to be useful in this task. However, when we apply the proposed confidence measure, the performance becomes worse and the WER is even lower than that without using any confidence measure. It is probably caused by the high WER. Since the proposed methods rely on the recognition results, high WER causes a poor estimation of confidence measure.

3. Domain-specific microphone speech recognition system

The removal of all incorrectly recognized data makes the WER even worse than the case of no confidence measure. It implies that some incorrectly recognized data are useful to adaptation but not all of them. It is observed that the performance is improved and close to that of supervised adaptation when the threshold is set to 0.1. Removing some data with very low CM really makes the adaptation more effective.

Selection of adaptation data is also required in this task but the definition of unreliable data may be different from that in recognition.

In our work, it is found that channel condition can affect the usefulness of the confidence measure for model adaptation. For the telephone channel recognition system, either general or specific domain, filtering the recognition error is useful to the adaptation. However, the negative result is observed in microphone speech. One of the reasons is that channel effect and phonetic mismatch both cause the recognition error in telephone speech while only the phonetic mismatch is involved in microphone speech. Such difference is one of the causes of discrepancy in the performance of confidence measure. The confidence measure can filter the error data due to the channel effect effectively but not the phonetic mismatch. Furthermore, the standard of so-called "bad" data to adaptation is different between two channels. Incorrectly recognized data may not be the bad adaptation data in microphone speech as well but it probably is in telephone speech.

Moreover, the word-density based confidence measure is based on the occurrence frequency in the N-best hypotheses. The baseline WER plays a deterministic role on whether the confidence score can be measured correctly. For a high baseline WER like in Task 3, either word or model level confidence measures cannot help the performance though the cheated confidence measure is found useful.

In conclusion, the removal of bad adaptation data is necessary in order to improve the adaptation. Confidence measure is helpful to certain extent to achieve the objective. However, the definition of bad and unreliable data in adaptation is sometimes different

to that in recognition. It is worth to continue the work on finding a suitable data selection method specifically for adaptation.

6.1. Future Works

1. Since the recognition error is not necessary to be bad to adaptation, it is necessary to add more information in the estimation. It is worth to further analyze the definition of bad adaptation to different speaking condition.

2. As we know that the not only correctness of the data in model level is important to the adaptation but also in state and class level, this information can be incorporated in the estimation.